

**Comparative phylogenetic exploration of
the human mitochondrial proteome:
Insights into disease and metabolism**

This dissertation is submitted for the degree of Doctor of Philosophy.

Cassandra Lauren Smith

Clare College

April 2018

Summary

Comparative phylogenetic exploration of the human mitochondrial proteome: insights into disease and metabolism

Cassandra Lauren Smith

Mitochondria are a key organelle within human cells, with functions ranging from ATP synthesis to apoptosis. Changes in mitochondrial function are associated with many diseases, as well as ‘natural’ processes like ageing. Mitochondria have a unique evolutionary origin, as the result of an endosymbiotic relationship between a bacterium and an archaeal cell. Therefore, the phylogenetic history of the mitochondrial proteome is also unique within the total human proteome. A new description of the genes encoding the human mitochondrial proteome – IMPI (Integrated Mitochondrial Protein Index) 2017 – provided an opportunity for exploration of mitochondrial proteome history and the application of this knowledge to the understanding of gene function, disease and ageing.

To facilitate the exploration of the mitochondrial proteome, I created a manually curated dataset of 190,097 predicted orthologues of the 1,550 IMPI 2017 human genes across 359 species, using reciprocal best hit analysis as the basis for orthologue prediction. I used this to explore gene history and the potential for phylogenetic profiling to predict the function of uncharacterised genes. This inspired the use of phylogenetic profiling within two phyla of animals, to link presence and absence of metabolic genes to the function of mitochondrial transporters. Potential transport substrates were predicted for two groups of uncharacterised mitochondrial carriers.

I also used the dataset to identify features of genes associated with monogenetic disease, as well as differences between recessive and dominant disease genes. A similar orthologue identification method was used to explore the total sequenced viral proteome for potential orthologues of mitochondrial proteins. This showed that a range of mitochondrial proteins are shared with viruses, potentially facilitating the co-opting of mitochondrial function during viral infection of eukaryotic cells. I then used orthology to explore the conservation of residues linked to protein acetylation and identify a link with lifespan in warm-blooded vertebrates.

In conclusion, I have used orthology to further the understanding of human mitochondrial proteome history and developed applications of this information. For example, phylogenetic features of disease genes are being used as part of a wider pipeline to predict mitochondrial disease genes. Furthermore, predicted substrates of the *SLC25A14/30* mitochondrial carriers are being tested. My dataset provides further opportunities to explore the evolution and function of the mitochondrion.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit of 60,000 words as designated by the School of Clinical Medicine Degree Committee.

Work in this thesis has been completed in collaboration with:

- Dr Alan Robinson of the Bioinformatics group, Medical Research Council Mitochondrial Biology Unit, University of Cambridge, CB2 0XY, UK (*Chapter 4*).
- Dr Anthony Smith of the Bioinformatics group, Medical Research Council Mitochondrial Biology Unit, University of Cambridge, CB2 0XY, UK (*Chapter 7*).
- Dr Andrew James of the Mitochondrial Dysfunction group, Medical Research Council Mitochondrial Biology Unit, University of Cambridge, CB2 0XY, UK (*Chapter 7*).

Acknowledgements

I would like to thank my supervisor Dr Alan Robinson for the chance to work in his group and for supporting me throughout my PhD. His questions and advice were always appreciated. Thanks to Dr Edmund Kunji for his guidance and enthusiasm on the topic of mitochondrial carriers. Thanks to Dr Mike Murphy and Dr Andrew James for the opportunity to be involved in their work on non-enzymatic lysine acetylation. Thanks, also, to Dr Anthony Smith for his work on the acetylation studies and for his advice on metabolism and other subjects. Thanks to the rest of the bioinformatics group and others in the lab for the chocolate, cakes and biscuits throughout my time here – always appreciated!

Thanks also to the Medical Research Council for funding my work throughout my PhD.

Thank you to all my friends and family for supporting me through this time and occasionally allowing me to forget that I had a thesis to write. Never again will you have to ask when I will get a ‘proper job’. You know who you are.

Contents

Title page	i
Summary	iii
Declaration	v
Collaborative work	vii
Acknowledgements	ix
Contents	xi
Chapter 1 Introduction	1
What are mitochondria?	3
Powerhouse of the cell	4
More than the powerhouse of the cell	6
The mitochondrial proteome	7
Mitochondrial evolution	8
The mitochondrial carrier family	9
Mitochondrial disease	10
Viruses and the mitochondria	11
Non-enzymatic lysine acetylation	11
The bioinformatic approach	12
Thesis outline	12
Chapter 2 Building an orthology dataset of genes encoding the human mitochondrial proteome	15
Introduction	17
<i>Homologues, orthologues and paralogues</i>	17
<i>Orthologue prediction</i>	19
<i>Chapter summary</i>	21
Methods	22
<i>Definition of the mitochondrial proteome</i>	22
<i>Choosing species and downloading proteomes</i>	22
<i>Reciprocal best hit analysis to predict orthologues</i>	23
<i>Identifying and utilising protein domain structure</i>	25
<i>Improving consistency of paralogue assignment</i>	25
<i>Phylogenetic tree</i>	26
<i>Gene enrichment</i>	26

Chapter 2	<i>Assignment of potential gene ancestry</i>	27
	Results & Discussion	28
	<i>Building an orthology dataset</i>	28
	<i>Using protein domains to improve orthologue predictions</i>	31
	<i>Manually improving paralogue assignment consistency</i>	33
	<i>Final orthology dataset</i>	39
	<i>Evolutionary history of the human mitochondrial proteome</i>	41
	Conclusions	48
Chapter 3	Investigating the mitochondrial respiratory complexes using phylogenetic profiling	49
	Introduction	51
	<i>Assembly of the mitochondrial respiratory chain complexes</i>	51
	<i>Phylogenetic profiling</i>	52
	<i>Chapter summary</i>	52
	Methods	53
	<i>Phylogenetic profiling of complex I</i>	53
	<i>Electron transfer flavoprotein phylogenetic profiling</i>	54
	<i>Complexes II-IV and ATP synthase phylogenetic profiling</i>	55
	Results & Discussion	56
	<i>Phylogenetic profiling of complex I: NADH dehydrogenase</i>	56
	<i>Electron transfer flavoprotein</i>	59
	<i>Other complexes and the limitations of phylogenetic profiling</i>	61
	Conclusions	67
Chapter 4	Function of mitochondrial carriers	69
	Introduction	71
	<i>Lessons from Chapter 3</i>	71
	<i>Mitochondrial carriers</i>	71
	<i>Chapter summary</i>	73
	Methods	74
	<i>Human carrier phylogenetic tree</i>	74
	<i>Identifying carriers by using sequence clustering and hidden Markov models</i>	74
	<i>Building nematode and platyhelminth orthologue datasets</i>	76
	<i>Clustering</i>	78
	<i>Pathway analysis</i>	78

Chapter 4	<i>Mitochondrial targeting sequence prediction</i>	79
	Results	80
	<i>Human mitochondrial carrier family</i>	80
	<i>Identifying useful species for phylogenetic profiling</i>	81
	<i>Building an orthologue dataset</i>	85
	<i>Investigating the characterised transporters</i>	90
	<i>Characterised transporters: SLC25A21</i>	90
	<i>Characterised transporters: SLC25A38</i>	94
	<i>Characterised transporters: SLC25A12/13</i>	101
	<i>Characterised transporters: SLC25A2/15</i>	104
	<i>Characterised transporters: SLC25A10</i>	106
	<i>Characterised transporters: SLC25A1 and SLC25A29</i>	107
	<i>Lessons from the characterised transporters</i>	108
	<i>Uncharacterised transporters</i>	109
	<i>Uncharacterised transporters: SLC25A14/30</i>	111
	<i>Uncharacterised transporters: SLC25A43</i>	118
	<i>Uncharacterised transporters: SLC25A44</i>	125
	<i>Uncharacterised transporters: SLC25A45/47/48</i>	130
	Discussion	132
	Conclusions	136
Chapter 5	Exploring the history and function of genes causing monogenetic mitochondrial diseases	137
	Introduction	139
	<i>Mitochondrial disease</i>	139
	<i>Features of disease genes</i>	140
	<i>Chapter summary</i>	141
	Methods	142
	<i>Defining a list of diseases of the mitochondrion</i>	142
	<i>Definitions of taxa</i>	142
	<i>Phylostratigraphic analysis</i>	143
	<i>Gene annotation and enrichment</i>	144
	<i>Essential genes in model organisms</i>	144
	<i>Human loss-of-function homozygotes</i>	145
	<i>Statistics</i>	146

Chapter 5	Results	147
	<i>Phylogenetic spread of monogenetic disease genes of the mitochondrion</i>	147
	<i>Phylogenetic origin of monogenetic disease genes of the mitochondrion</i>	150
	<i>Inheritance patterns of IMPI genes associated with monogenetic mitochondrial disease</i>	155
	<i>Functional analysis</i>	160
	<i>Human loss-of-function (LoF) homozygotes</i>	163
	<i>Essential genes</i>	167
	Discussion	173
	Conclusions	176
Chapter 6	Mitochondrial proteins in viruses	177
	Introduction	179
	<i>Viruses and the mitochondria</i>	179
	<i>Viral orthologues of mitochondrial proteins</i>	180
	<i>Chapter summary</i>	180
	Methods	182
	<i>Identifying viral orthologues of human mitochondrial genes</i>	182
	<i>Information on viruses</i>	183
	<i>Functional enrichment</i>	183
	<i>Phageness</i>	183
	<i>Matrix clustering</i>	184
	<i>Localisation</i>	184
	<i>Gene families</i>	184
	<i>MFTP1 phylogenetic tree</i>	184
	<i>Viral mitochondrial carrier analysis</i>	185
	<i>Statistics</i>	186
	Results	187
	<i>Identifying orthologues of human mitochondrial genes in viruses</i>	187
	<i>Viruses with predicted mitochondrial gene orthologues</i>	189
	<i>Function of viral orthologues of mitochondrial proteins</i>	192
	<i>Spread of viral orthologues of mitochondrial proteins</i>	196
	<i>Transfer of genes between virus and host genomes</i>	198
	<i>Mitochondrial carriers in viruses</i>	201

Chapter 6	Discussion	205
	Conclusions	207
Chapter 7	Conservation of non-enzymatic $S \Rightarrow N$ lysine acetylation	209
	Introduction	211
	<i>Lysine N-acetylation</i>	211
	<i>Animal longevity</i>	212
	<i>Chapter summary</i>	213
	Methods	214
	<i>Identifying, close surface pairs</i>	214
	<i>Identifying orthologues of acetylated mouse proteins</i>	214
	<i>Estimating relative conservation of residues and pairs</i>	217
	<i>Defining cytosolic and mitochondrial matrix proteins</i>	219
	<i>Estimating lifespan of sequenced vertebrates</i>	219
	<i>Statistics</i>	220
	Results	221
	<i>Creating a dataset of cysteine-lysine pair conservation</i>	221
	<i>Close acetylated Cys-Lys pairs are less conserved in the cytosol</i>	222
	<i>Conservation of cytosolic Cys-Lys pairs correlates with a measure of mammalian lifespan</i>	226
	Discussion	231
	Conclusions	234
Chapter 8	Conclusions	235
	Thesis summary	237
	Further work	239
	Final words	241
	References	243
	Appendices	281
	Appendix I: Orthology dataset summary	283
	Appendix II: Mitochondrial genes matching mitochondrial carrier phylogenetic patterns in nematodes and platyhelminthes	311
	Appendix III: Mitochondrial proteins in viruses	319
	Appendix IV: Acetylation study species	326

Chapter 1

Introduction

What are mitochondria?

Mitochondria are double-membrane bound organelles found in nearly all eukaryotic cells.

Figure 1.1 shows a traditional diagram of a single mitochondrion, with the two membranes and the folds of the inner membrane known as cristae. Electron tomography work has shown that the cristae are only connected to the boundary inner mitochondrial membrane by small structures known as crista junctions, which allows tight regulation of protein and metabolite distribution within the cristae (Zick *et al.* 2009).

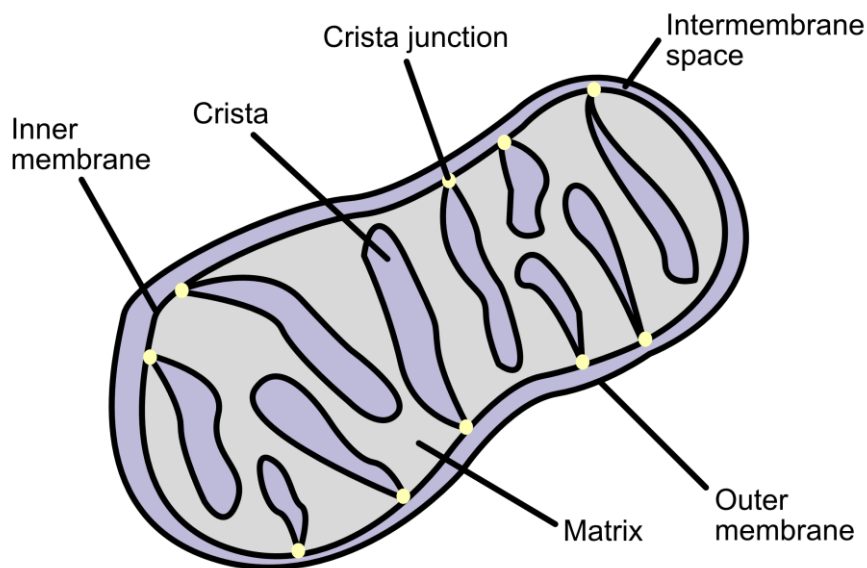


Figure 1.1 Cut-through structure of a single mitochondrion, showing the inner and outer mitochondrial membranes, cristae and crista junctions (yellow dots).

In vivo mitochondrial structure is more complex. Mitochondria are dynamic organelles in a constant state of flux, with fusion and fission facilitating the creation of an interconnected network (Chan 2006). Control of this network is important for maintaining mitochondrial and cellular health, and dysfunction of the network has been implicated in both monogenetic disease and neurodegenerative disorders (Westermann 2010). Mitochondria also form connections with other parts of the cell, including the endoplasmic reticulum (Giacomello & Pellegrini 2016), making them a highly connected and dynamic part of eukaryotic cells.

Powerhouse of the cell

Mitochondria are often described as the ‘powerhouse of the cell’ and, indeed, they are the main site of aerobic production of adenosine triphosphate (ATP) – the energetic currency – in most eukaryotic cells. In mammals, including humans, the majority of ATP is produced by ATP synthase, located in the inner mitochondrial membranes which form cristae (Fernandez-Moran 1962). ATP synthase activity is dependent on the proton gradient across the membrane (ΔpH), as well as the charge gradient across the membrane ($\Delta\psi$), which combined are known as the proton motive force.

In humans, the proton differential is produced by the action of the four protein complexes (complexes I – IV) which form the electron transport chain (*Figure 1.2*), plus the electron transfer flavoprotein (ETF). In general, these complexes couple energy-releasing redox reactions, involving electron transport between electron donors of increasing redox potential, to the transport of protons across the inner mitochondrial membrane into the intermembrane space formed by the cristae. Electrons can enter the electron transport chain from two electron carriers: NADH (nicotinamide adenine dinucleotide) and FADH_2 (flavin adenine dinucleotide).

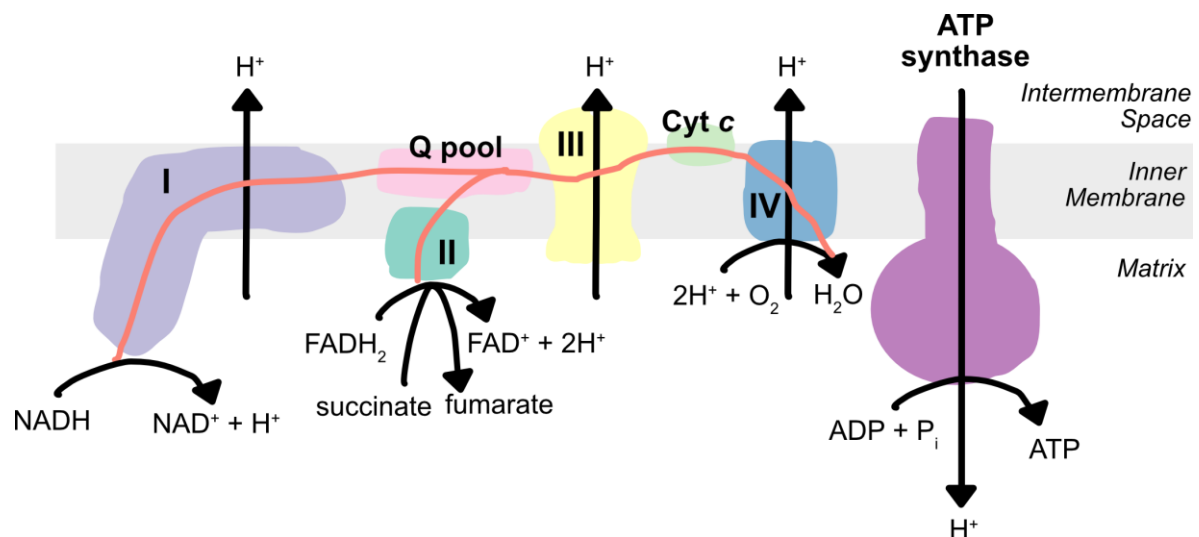


Figure 1.2 Summary of the complexes of the electron transport chain and ATP synthase. The red line shows the flow of electrons. *Q pool* = pool of ubiquinol/ubiquinone; *Cyt c* = cytochrome c.

Electrons from NADH enter the electron transport chain at complex I (NADH dehydrogenase), transferring to the flavin mononucleotide prosthetic group (Zhu *et al.* 2016). Electrons then move down a series of iron-sulphur clusters of increasing redox potential,

which is coupled to the movement of four protons across the inner mitochondrial membrane. Finally, the electrons, plus additional protons, are transferred to the hydrophobic electron carrier ubiquinone, forming ubiquinol which is free to move in the inner mitochondrial membrane.

Electrons from FADH_2 enter the electron transport chain at either complex II (succinate dehydrogenase) (Sun *et al.* 2005) or the ETF (Roberts *et al.* 1996). The reactions catalysed by these complexes are not directly coupled to proton transport. FAD^+ is covalently bound to these complexes. Complex II is also part of the tricarboxylic acid cycle. Electrons and protons are released from the oxidation of succinate to fumarate and transferred onto a covalently attached FAD^+ cofactor on complex II, forming FADH_2 . These electrons, along with two protons, are eventually transferred onto ubiquinone forming ubiquinol – the same endpoint as with complex I. Electrons from ETF also enter the ubiquinol pool, via the action of ETF dehydrogenase.

Electrons from complex I, complex II and ETF, therefore, converge in the ubiquinol pool, and move through the remaining complexes in the same way. Reduced ubiquinol delivers electrons to complex III (coenzyme Q : cytochrome *c* oxidoreductase), with four protons pumped to the intermembrane space per ubiquinol (Rieske 1976). The reactions of complex III result in the production of two reduced molecules of cytochrome *c*, which move to complex IV (cytochrome *c* oxidase) (Capaldi 1990). Electrons moving through complex IV are used to reduce oxygen to water, with four protons pumped to the intermembrane space per oxygen molecule reduced, as well as using four matrix protons to form the water molecules.

ATP is produced by the controlled re-entry of protons to the mitochondrial matrix through ATP synthase, driving the rotation of the *c*-ring. This is coupled to the production of ATP from ADP and inorganic phosphate (P_i) (Yoshida *et al.* 2001). As animals have eight *c*-subunits in their *c*-ring rotor and three ATPase subunits in the hydrophilic F_1 domain, the movement of approximately 2.7 protons through ATP synthase into the mitochondrial matrix produces one molecule of ATP (Watt *et al.* 2010). ATP can then either be used within the mitochondria or transported out to the cytosol through the ATP/ADP transporter (Kunji 2004), to act as a source of energy in a variety of biological processes.

More than the powerhouse of the cell

The electron transport chain and ATP synthesis are only a small part of mammalian mitochondrial function. Mitochondria are also the site of a variety of metabolic pathways, as well as being linked to a range of other cellular processes (McBride *et al.* 2006). Central mitochondrial metabolism includes pathways with obvious links to energy metabolism, such as the tricarboxylic acid cycle, fatty acid β -oxidation and the synthesis/consumption of ketone bodies; but also steps of nucleotide metabolism, amino acid metabolism (Nunnari & Suomalainen 2012) and steroid synthesis (Miller 2013).

Mitochondria are important in the synthesis and processing of several cofactors. They are the location of the iron-sulphur cluster (ISC) assembly pathway (Stehling & Lill 2013) and a part of the haem synthesis pathway (Ryter & Tyrrell 2000) – both cofactors which coordinate iron. Mitochondria are also the site of the formation of adenosylcobalamin (an active derivative of vitamin B12) (Quadros 2010) and one of the early steps in the synthesis of molybdenum cofactor, which coordinates the transition metal molybdenum (Schwarz *et al.* 2009). These cofactors can be utilised by proteins within the mitochondria and/or other sub-compartments of the cell.

Beyond metabolism and biosynthesis, mitochondria are involved in the process of apoptosis – the programmed death of a cell (Wang & Youle 2009). In particular, they participate in the intrinsic pathway of apoptosis (a.k.a. the mitochondrial pathway) which initiates cell death when the cell suffers certain stresses (e.g. DNA damage). These stresses trigger the permeabilisation of the mitochondrial membranes, causing the release of mitochondrial proteins into the cytoplasm. Several of these proteins have defined roles in apoptosis. For example, release of cytochrome *c* into the cytoplasm triggers the formation of the apoptosome (Zou *et al.* 1999), which drives the process of cell death.

The mitochondrial apoptotic pathway is also connected to their involvement in calcium ion (Ca^{2+}) homeostasis and signalling (Contreras *et al.* 2010). The mitochondrial Ca^{2+} concentration is very responsive to the cellular Ca^{2+} concentration, despite the low affinity of the mitochondrial calcium uniporter for Ca^{2+} (Rizzuto *et al.* 1993). This is due to contact sites between the endoplasmic reticulum (a Ca^{2+} storage site) and mitochondrial membranes, exposing mitochondria to localised high concentrations of Ca^{2+} (Csordás *et al.* 2006). Mitochondria act as overflow Ca^{2+} storage sites. Large and sustained increases in

mitochondrial Ca^{2+} concentration have been linked to the permeabilisation of mitochondrial membranes and, thus, apoptosis (Contreras *et al.* 2010).

More complex cell-wide processes, such as movement through the cell cycle and signals controlling gene expression, have also been linked to the mitochondria (McBride *et al.* 2006), giving the mitochondria roles in both large- and small-scale cellular functions.

The mitochondrial proteome

The mitochondrial proteome consists of the proteins which localise to the mitochondria. These proteins then function in the many parts of metabolism and cellular activity in which mitochondria have been implicated. The size of the mammalian mitochondrial proteome has been predicted to be around 1,500 proteins (Meisinger *et al.* 2008). Human mitochondria contain their own circular genome (mtDNA). However, human mtDNA only encodes thirteen proteins (membrane subunits of the electron transfer chain complexes), twenty-two transfer RNAs and two ribosomal RNAs (Anderson *et al.* 1981). It is clear that the majority of the human mitochondrial proteome must be encoded on the nuclear genome and imported into the mitochondria, but which of the proteins encoded by the nuclear genome are part of the mitochondrial proteome?

There have been several attempts to predict the human mitochondrial proteome, but there is not a definitive description. MitoP2 (Elstner *et al.*, 2008) and MitoProteome (Guda *et al.* 2007) are two older efforts, but these are no longer available online. MitoCarta (Pagliarini *et al.* 2008) and the more recent update MitoCarta 2.0 (Calvo *et al.* 2015) are two well-used predictions of the human mitochondrial proteome. MitoCarta integrates data from seven different experimental methodologies using a Bayesian classifier, producing a dataset of 1,158 predicted human genes encoding the mitochondrial proteome.

IMPI (Integrated Mitochondrial Protein Index) 2017 is a new attempt to predict the mitochondrial proteome (<http://www.mrc-mbu.cam.ac.uk/impi>). IMPI was created by training a machine learning classifier to separate training sets of mitochondrial and non-mitochondrial proteins based on a wide range of evidence types, some of which have not been previously used in attempts to define the mitochondrial proteome. These include: a large range of mass spectrometry studies, GFP (green fluorescent protein) tagging studies, mitochondrial targeting

sequence predictions from four programs and several species, antibody data and more. This classifier was then used to score all human genes based on their evidential properties, with genes above a score threshold of 0.8 assigned to the human mitochondrial proteome. The human mitochondrial proteome defined by IMPI 2017 consists of proteins encoded by 1,550 genes: 1,130 from the mitochondrial positive training set and 420 predictions. This more extensive proteome prediction provides an excellent source for investigating the history of the human mitochondrial proteome.

Mitochondrial evolution

Current evidence supports the ‘endosymbiotic theory’ of the origin of mitochondria, made famous by Lynn Margulis in 1967 (Sagan 1967), although the idea had been discussed much earlier in the century by Ivan Wallin (Wallin 1927). This theory proposes that mitochondria are the result of an ancient, mutually beneficial relationship between free-living bacteria and a host cell which somehow took up these bacteria. Evidence for this theory includes the double cellular membrane around the mitochondria (one from the original bacterium and one derived from the host on take-up of the bacterium); the presence and structure of the mitochondrial genome, which is circular as in bacteria (Anderson *et al.* 1981); and the independent reproduction of mitochondria by binary fission, as seen in bacteria (Margolin 2005). Phylogenetic analysis has suggested that the bacteria which eventually became the mitochondria were members of the α -proteobacteria, related to the order Rickettsiales (Andersson *et al.* 1998) whereas the host cell was potentially archaeal, though this is still under discussion (Gribaldo *et al.* 2010).

There are two general groups of hypotheses about how this uptake may have occurred (Gray 2015). The first group suggest that uptake happened by phagocytosis of the bacteria by a host cell which already had some eukaryotic features, where the bacteria then somehow survived and became functionally useful to the host. The second suggests that the bacteria and host developed some form of metabolic dependence as free-living organisms. The bacteria were then eventually taken up by the host, which sparked the development of cellular complexity observed in eukaryotic cells. This second hypothesis is supported by theoretical work which suggests the genome complexity required by eukaryotes was supported by the huge increase in bioenergetic membrane surface area that mitochondria provide (Lane & Martin 2010).

While mitochondria are only found in eukaryotic species, not all eukaryotic species have mitochondria. One early explanation for this was the Archezoa theory (Cavalier-Smith 1989). This theory suggested that there were primitive eukaryotic organisms, known as Archezoa, which had developed much of the complexity of eukaryotic cells (nucleus, endoplasmic reticulum etc.) but not mitochondria. However, later work showed that these species thought to lack mitochondria instead have degenerate forms of mitochondria, which vary in size and function. Examples include mitosomes in the Microsporidia, which retain iron-sulphur cluster assembly machinery (Goldberg *et al.* 2008); mitosomes in *Entamoeba* species, which retain a sulphate activation pathway (Mi-Ichi *et al.* 2009); and hydrogenosomes in *Trichomonas vaginalis*, the site of an anaerobic form of carbon metabolism (Müller 1993) and iron-sulphur cluster assembly (Sutak *et al.* 2004; Dolezal *et al.* 2007). Recently, an oxymonad (*Monocercomonoides* sp.) has been sequenced with no recognisable mitochondrial marker proteins, suggesting possible loss of the entire mitochondrial organelle (Karnkowska *et al.* 2016).

The remarkable history of mitochondria and the wide variation in function across species make the genes encoding the mitochondrial proteome a particularly interesting group for phylogenetic-based analysis.

The mitochondrial carrier family

One potentially interesting group of proteins is the mitochondrial carrier family. In general, mitochondrial carrier proteins are localised to the mitochondrial inner membrane, and act to move compounds or ions in either direction across this membrane (Gutiérrez-Aguilar & Baines 2013). Transporter family proteins are necessary as, unlike the outer mitochondrial membrane, the inner mitochondrial membrane is relatively impermeable. This allows strict control of the molecules which can enter and leave the mitochondrial matrix, creating a compartmentalised environment within the cell.

There are fifty-three known members of the mitochondrial carrier family in humans (Gutiérrez-Aguilar & Baines 2013). Several of these have characterised transport substrates. For example, the four carriers encoded by the genes *SLC25A4*, *SLC25A5*, *SLC25A6* and *SLC25A31* are known to exchange mitochondrial ATP for cytosolic ADP (Dolce *et al.* 2005). Mitochondrial carrier proteins are also known to transport a variety of metabolites and ions.

Other mitochondrial carriers have been associated with particular pathways, even where the exact transport substrate remains unknown. For example, the mitochondrial carrier encoded by *SLC25A38* has been associated with haem biosynthesis (Guernsey *et al.* 2009). It has been suggested that this is due to its function as a glycine transporter, as glycine is an early substrate for haem biosynthesis (Guernsey *et al.* 2009). This has not been biochemically confirmed using transport assays.

However, there are several mitochondrial carriers with no associated transport substrate, pathway or function (Gutiérrez-Aguilar & Baines 2013). Understanding the function of these carriers would increase our knowledge of the working of human mitochondria and potential associated diseases.

Mitochondrial disease

Mitochondrial dysfunction has been implicated in many types of disease, ranging from late-onset, complex diseases, such as neurological disorders (Onyango *et al.* 2017), to paediatric-onset, monogenetic diseases, such as those involving the respiratory chain (DiMauro & Schon 2003). Currently, exome sequencing produces genetic diagnoses for approximately half of the studied patients assumed to be suffering from mitochondrial disease caused by mutations in single genes (Calvo *et al.* 2012; Taylor *et al.* 2014). One bottleneck for identifying new genes associated with mitochondrial disease is prioritising the results from sequencing (which identifies variants in a number of different genes) for biological study.

Learning more about the features of known mitochondrial disease genes may help in the prioritisation of candidates for new disease genes in future mitochondrial disease patients. Understanding differences between disease genes associated with different types of disease inheritance (recessive versus dominant) could also be informative. This is important to consider as evidence is beginning to support a larger contribution of dominant *de novo* mutations to disease burden than may have been expected. For example, around 40% of a sequenced developmental disease cohort encoded *de novo* variants that were predicted to be pathogenic (McRae *et al.* 2017). Understanding the phylogenetic properties of these disease-associated genes may, therefore, be of interest.

Viruses and the mitochondria

Mitochondria, and the various functions they are involved in, have been implicated in the lifecycle of a variety of viruses (Ohta & Nishiyama 2011). This may include viral hijacking or processes such as metabolism, apoptosis, DNA replication, transcription and translation. Viruses encode a range of proteins to aid in the manipulation and control of host cell mitochondria. These proteins can be novel to viruses, or proteins which appear orthologous to eukaryotic proteins known to localise to the mitochondria and/or affect mitochondrial function (Ohta & Nishiyama 2011). For example, a variety of Bcl-2 family proteins (associated with the mitochondrial pathway of apoptosis in humans (Gross *et al.* 1999)) have been identified in viruses (Cuconati & White 2002). While there are many individual studies studying orthologues of mitochondrial proteins in viruses, a wider scale study looking at the total known viral proteome has not been carried out, leaving an opening for research into this area.

Non-enzymatic lysine acetylation

Proteins throughout the cell, including those localised to mitochondria, can be modified by the addition of a range of chemical groups. Modifications include the addition of an acetyl group to the ϵ -amino group of exposed lysine residues on a protein surface, which is known as lysine *N*-acetylation. Though acetylation is often thought of as a controlled process which is important for protein regulation, non-enzymatic protein acetylation occurs in conditions replicating those seen in the mitochondrial matrix (Wagner & Payne 2013). More recent work has shown that acetyl groups can be transferred non-enzymatically from acetyl-CoA to a protein lysine residue, via a nearby cysteine residue (James *et al.* 2017).

It has been hypothesised that this non-enzymatic acetylation is non-functional and potentially damaging to the cell – a ‘carbon stress’ (Wagner & Hirschey 2014). The idea that general non-enzymatic acetylation may be damaging is of particular interest as changes in acetylation have been linked to changes in lifespan in several model organisms, including yeast and mice, particularly in relation to caloric restriction (Lin *et al.* 2000; Schwer *et al.* 2009). Though these experiments have thought about the link between non-enzymatic acetylation and

lifespan within species, it may be of interest to consider acetylation as a modifier of lifespan between species.

The bioinformatic approach

The work in this thesis uses bioinformatic techniques in a variety of ways to investigate the mitochondrial proteome, including looking at protein function, features of disease genes and conservation of acetylation motifs. Bioinformatics involves the development and use of computational techniques to organise and analyse data (Rhee 2005). Bioinformatic approaches, in general, are at their best as a set of methods used to make predictions and prioritise hypotheses, guiding the development of wet laboratory experiments. For example, bioinformatic methods can be used to predict the association of a protein with a complex or pathway. This could then guide biological experiments which focus around function of the complex or pathway of interest, rather than looking more generally at the function of the cell or organism.

Bioinformatic techniques are also useful to study topics which are currently experimentally intractable. This could include experiments which would need to occur over long periods of time, those that would be prohibitively expensive, or those that require a degree of experimental specificity not currently available to scientists. In these cases, bioinformatic methods can provide information to support a hypothesis, though experimental confirmation is preferable where this becomes possible.

In summary, the bioinformatic approaches inform the biological approaches but are usually not an endpoint in themselves.

Thesis outline

This thesis describes the use of a newly predicted human mitochondrial proteome, defined by IMPI 2017, as the basis to explore the history of the mitochondrial proteome from a human perspective; as well as several applications of this exploration in aspects of disease, metabolism and function of the mitochondria.

In *Chapter 2*, I describe the creation of an orthology dataset of the genes encoding the human mitochondrial proteome, over a large range of species (both eukaryotic and prokaryotic). I use reciprocal best hit analysis as a basis for predicting orthology. From the resulting dataset, I explore the potential development of the human mitochondrial proteome over time.

In *Chapter 3*, I present the use of this orthology dataset to predict genes associated with the function of the respiratory complexes, particularly complex I, using the predicted presence and absence of genes across the investigated species (the phylogenetic profile).

In *Chapter 4*, I use a dataset of mitochondrial proteome orthologues in two metazoan phyla (Nematoda and Platyhelmintha) to predict the transport substrates of uncharacterised members of the mitochondrial carrier family, using phylogenetic profiling and extensive manual assessment of the resulting patterns.

In *Chapter 5*, I use the orthology dataset to explore the genes of the mitochondrial proteome in terms of their association with monogenetic diseases, including a comparison of recessive and dominant disease genes. I also assess the utility of model organisms in the prediction of disease genes.

In *Chapter 6*, I describe the prediction of potential viral orthologues of human IMPI 2017 genes. I explore some general features of the genes and viruses with predicted orthologues and look further into some groups of predicted orthologues.

In *Chapter 7*, I describe an analysis of the conservation of sites of non-enzymatic acetylation in cytoplasmic and mitochondrial proteins across vertebrates; and a potential link to the lifespan of these species.

In *Chapter 8*, I present a brief summary of the findings from each chapter, as well as a discussion of further work which may be prompted by this thesis.

Chapter 2

Building an orthology dataset of genes
encoding the human mitochondrial
proteome

Introduction

Homologues, orthologues and paralogues

There are three key classifications for related genes (Fitch 2000). Genes which are orthologous to each other (orthologues) are genes which have been separated by a speciation event – in *Figure 2.1*, Gene 1 from Species 1 and Gene 1 from Species 2 are orthologues. Genes which are paralogous to each other (paralogues) are genes which have been separated by a duplication event. Genes can be in-paralogues – two genes in the same species separated by a duplication event – as with Gene 1 and Gene 2 from Species 1 in *Figure 2.1*. Genes can also be out-paralogues – two genes in different species, separated by both a speciation event and a duplication event – as with Gene 2 from Species 1 and Gene 1 from Species 2 in *Figure 2.1*. Genes which are homologous to each other (homologues) may have been separated by speciation, duplication or a combination of both events – all genes in *Figure 2.1* are homologues.

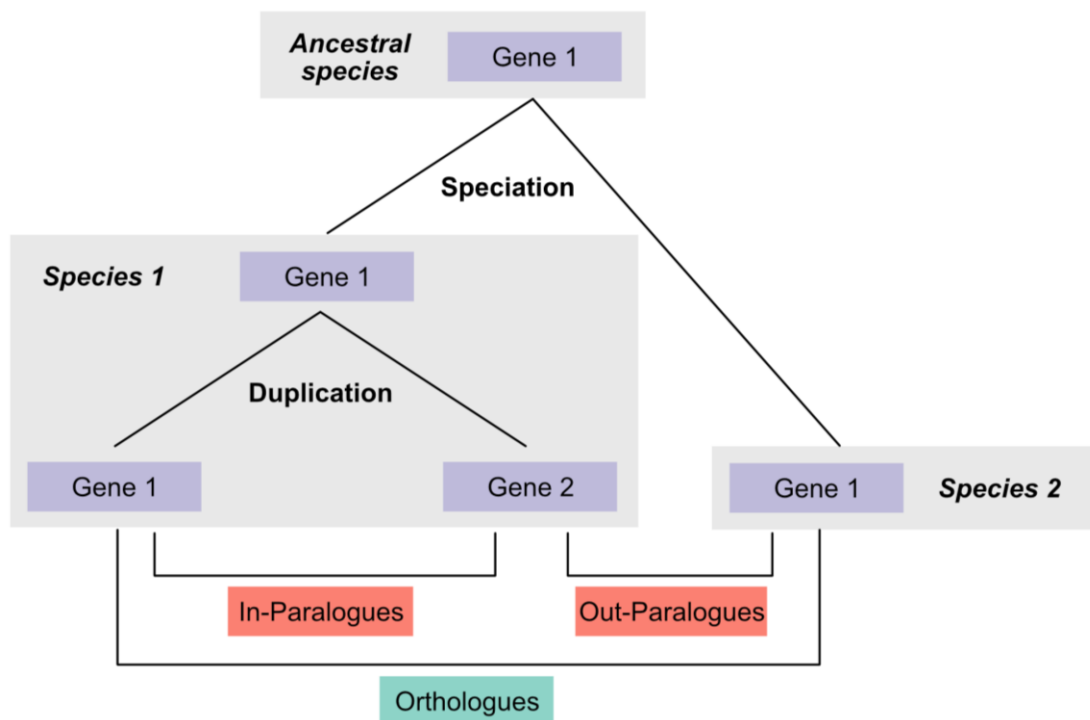


Figure 2.1 Examples of orthologues and paralogues of a single ancestral gene, after speciation and duplication events.

One assumption that is often made is that orthologues will encode proteins with more similar functions than paralogues (Tatusov *et al.* 1997; Koonin 2005a). The theory behind this is that a duplication event produces a paralogous gene which is not under the same evolutionary pressure to conserve function as the original gene, as the function is maintained by the original gene. Therefore, the paralogous gene can diverge in both sequence and function (Conant & Wolfe, 2008).

Testing this assumption is complicated as, in many cases, gene function has not been experimentally confirmed across a range of species, and functional annotation may have been propagated over predicted orthologous sequences automatically. Different aspects of gene function may also have been studied in different species, biasing the annotation. One study, comparing the functional annotation of genes in humans and mice, seemed to refute the traditional view of homologous function, finding that paralogues were more functionally similar than orthologues (Nehrt *et al.* 2011). However, the methodology used in this study has been challenged (Thomas *et al.* 2012) and a later study, looking at a larger number of species and correcting for more potential biases, found that orthologues were significantly more similar in function than paralogues, though the difference was smaller than may have been expected (Altenhoff *et al.* 2012). This supports the idea of trying to separate orthologous and paralogues when investigating homologous sequences, where conservation of function is often assumed.

Whilst conserved function is often assumed when predicting orthologues, computational orthology prediction methods do not usually take function into account, instead looking at sequence similarity and/or positions in phylogenetic trees. This is important to keep in mind when using predicted orthologues in analyses in which conservation of function is an important assumption – for example, when trying to predict the function of an uncharacterised gene using phylogenetic profiling, as is discussed later in this thesis (*Chapter 3 & Chapter 4*).

Orthologue prediction

Understanding the history of the human mitochondrial proteome requires two things. The first is a good definition of the genes which encode the mitochondrial proteome, which is now provided by IMPI 2017. The second is a set of orthologous sequences of these genes from species spread throughout the tree of life.

Many methods have been developed that attempt to predict orthologous sequences. Hulsen *et al.* (2006) suggest that the method of choice should depend on the question being asked – what are the performance measures that are most important to you? This was supported by a more wide-scale comparison of the sensitivity (a measure of the proportion of false negatives) and specificity (a measure of the proportion of false positives) of different orthology detection methods which suggested improvement in one performance measure was often a trade-off with lower performance in the other (Altenhoff & Dessimoz 2009). This study also suggested one of the simpler methods of orthology detection (best bidirectional hit or reciprocal best hit) performed well when compared to more complex methodologies (Altenhoff & Dessimoz 2009). A more recent comparison of fifteen common orthologue detection methodologies, using a suite of standardised benchmarking tests, confirmed that reciprocal best hit was one of the most specific of the tested methodologies, with an average level of sensitivity (Altenhoff *et al.* 2016). This is consistent with the results of other smaller-scale analyses comparing different methods of orthology detection (Chen *et al.* 2007; Dalquen & Dessimoz 2013).

The reciprocal best hit method uses two separate searches – in this case BLASTp (Basic Local Alignment Search Tool) searches (Altschul *et al.* 1990, 1997) – to predict orthologues and reduce the identification of paralogues (Rivera *et al.* 1998). To be deemed orthologous, sequences from two different species must be the best hit of each other, when searching across the opposite species' genome (*Figure 2.2*).

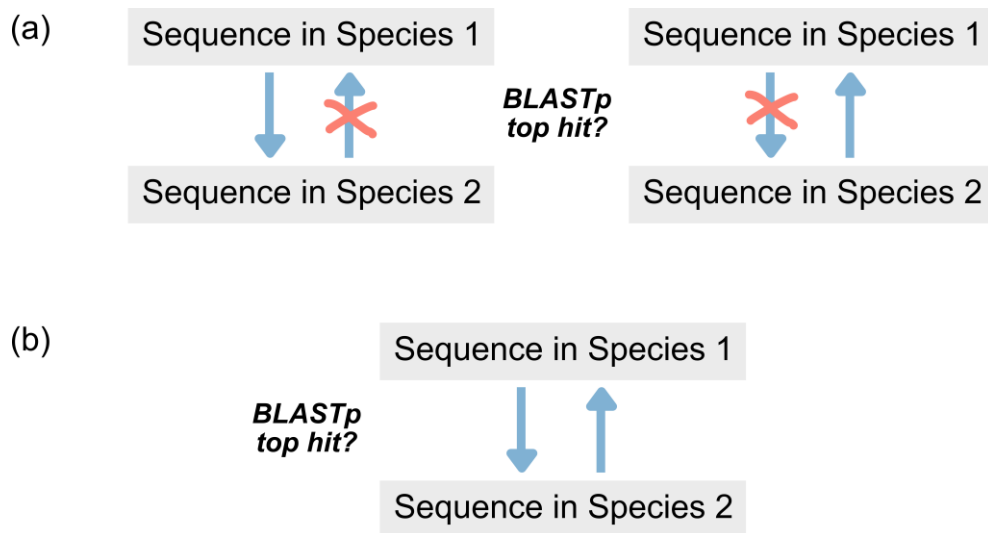


Figure 2.2 Principle of reciprocal best hit orthologue identification. (a) Examples of failures to identify an orthologue. A BLASTp search using the species 1 sequence returns a sequence in species 2 as the top hit. However, the search back using the species 2 sequence does not return the original species 1 sequence, or vice versa. The sequences from each species are not the top hit of the other. (b) Example which would return a pair of predicted orthologues, as the sequences from each species are the top hit of the other, using BLASTp searches.

The aim of this work was to create a history of orthologous sequences in relation to human genes, not to identify all homologues within the studied species that form a gene family. Reciprocal best hit can be centred around the genes of a particular species. So, in this case, genes can be specifically identified as orthologues of human genes, rather than as groups of homologous genes from a gene family. This should facilitate good separation of paralogues in species closely related to humans. Reciprocal best hit is also relatively easy to set up and run for a range of species, which can be defined by the user. This was important for some of the potential uses for the dataset (e.g. to investigate the presence and absence of specific genes in certain species of interest). Therefore, I based the creation of a dataset of orthologues of genes encoding the human mitochondrial proteome on reciprocal best hit analysis.

Chapter summary

In this chapter, I describe the automated creation and manually curated improvement of a set of predicted orthologues of genes encoding the human mitochondrial proteome, as defined by IMPI 2017, across a variety of eukaryotic and prokaryotic species. A summary of the properties of this dataset is provided, including an analysis of the potential origins of the genes encoding the human mitochondrial proteome.

Methods

Definition of the human mitochondrial proteome

The genes encoding the human proteome were defined as the genes included in the 2017 release of IMPI (Integrated Mitochondrial Proteome Index) developed by the Bioinformatics group of the Medical Research Council Mitochondrial Biology Unit (available from www.mrc-mbu.cam.ac.uk/impi). IMPI was built from a number of both publicly available and custom evidence sources looking at the mitochondrial localisation of proteins, including network analysis scores, mitochondrial targeting prediction programs, mass spectrometry analyses and more. This evidence was provided as input to a supervised machine learning classifier, along with manually curated lists of positive training genes (for proteins known to localise to mitochondria) and negative training genes (for proteins believed not to localise to mitochondria), and the classifier used these data to identify the most important evidential properties of these training sets. The classifier was then used to assign each human gene a score from between 0 and 1, that predicted its likelihood of being mitochondrial. Genes from the positive training set and other genes with a score of 0.8 or above were included in IMPI 2017, producing a list of 1,550 human genes predicted to encode the mitochondrial proteome.

Choosing species and downloading proteomes

A selection of species was chosen from those with fully sequenced genomes and predicted proteomes in the NCBI Protein database (www.ncbi.nlm.nih.gov/protein). Species were selected to cover a representative range from the eukaryotes, archaea and bacteria, as well as to cover species with known variations in the function of their mitochondria. The internet and literature were also searched for sites and papers describing sequenced species whose taxa are not covered in the NCBI database, and which had publicly available predicted proteomes (Table 2.1). This resulted in the choice of 359 species for orthology analysis (see Appendix I – Table 1).

Table 2.1 Additional sources of species' proteomes, for those not retrieved from the NCBI Protein database.

Species	Download source	Reference
<i>Schmidtea mediterranea</i>	http://parasite.wormbase.org/	1
<i>Mnemiopsis leidyi</i>	http://research.nhgri.nih.gov/mnemiopsis/	2
<i>Symbiodinium minutum</i> Clade B1	http://marinegenomics.oist.jp/gallery/	3
<i>Cyanophora paradoxa</i>	http://cyanophora.rutgers.edu/cyanophora/	4
<i>Monocercomonoides</i> sp. PA203	Supplementary information of reference	5

1 – (Robb *et al.* 2015); 2 – (Ryan *et al.* 2013); 3 – (Shoguchi *et al.* 2013); 4 – (Price *et al.* 2012); 5 – (Karnkowska *et al.* 2016)

For each species' proteome, each sequence was checked for an individual identifier – paying particular attention to the sequences from outside the NCBI Protein database. Sequences which were completely identical (i.e. redundant) within each species' proteome were removed.

Reciprocal best hit analysis to predict orthologues

First, a custom BLAST database was created, combining the non-redundant proteomes of each of the 359 species included in the analysis. The canonical human protein sequence for each of the 1,550 IMPI 2017 genes was identified and downloaded from the NCBI Protein database. The sequence with 'RecName' (a UniProt assignment) in the title was chosen as the canonical sequence where possible. Else, the longest protein sequence associated with the gene was chosen.

For each of these human protein sequences, a local BLASTp search (Altschul *et al.* 1990, 1997) was carried out against the custom BLAST database, with default settings, and an Expect-value (E-value) threshold of 10. For each of the 359 species, the top protein hit with the smallest E-value was identified. If two or more proteins in one species returned the same E-value, the sequence with the highest bit score was chosen. If two or more sequences also returned the same bit score, both sequences were taken forward (Moreno-Hagelsieb & Latimer 2008).

A second custom BLAST database was then created. This database contained all human protein sequences (NCBI: txid9606) that could be associated via their annotated gene name

with a list of all human gene names downloaded from the HGNC (HUGO Gene Nomenclature Committee) (Gray *et al.* 2014).

A BLASTp search (Altschul *et al.* 1990, 1997) was then carried out against this human protein database, for each of the top protein hits, for each species, for each IMPI 2017 gene. The results of these reciprocal BLAST searches were analysed to predict orthologues. The top hits from the reciprocal search were identified as in the original search – the sequence with the lowest E-value; then the sequence with the highest bit score. All sequences that fitted these criteria were considered.

A sequence was declared a potential orthologue if the top result was a protein sequence encoded by the same human gene that the analysis started with. If the top result was a human protein sequence encoded by any other human gene, the sequence was not declared an orthologue.

This reciprocal best hit process is summarised in *Figure 2.3*.

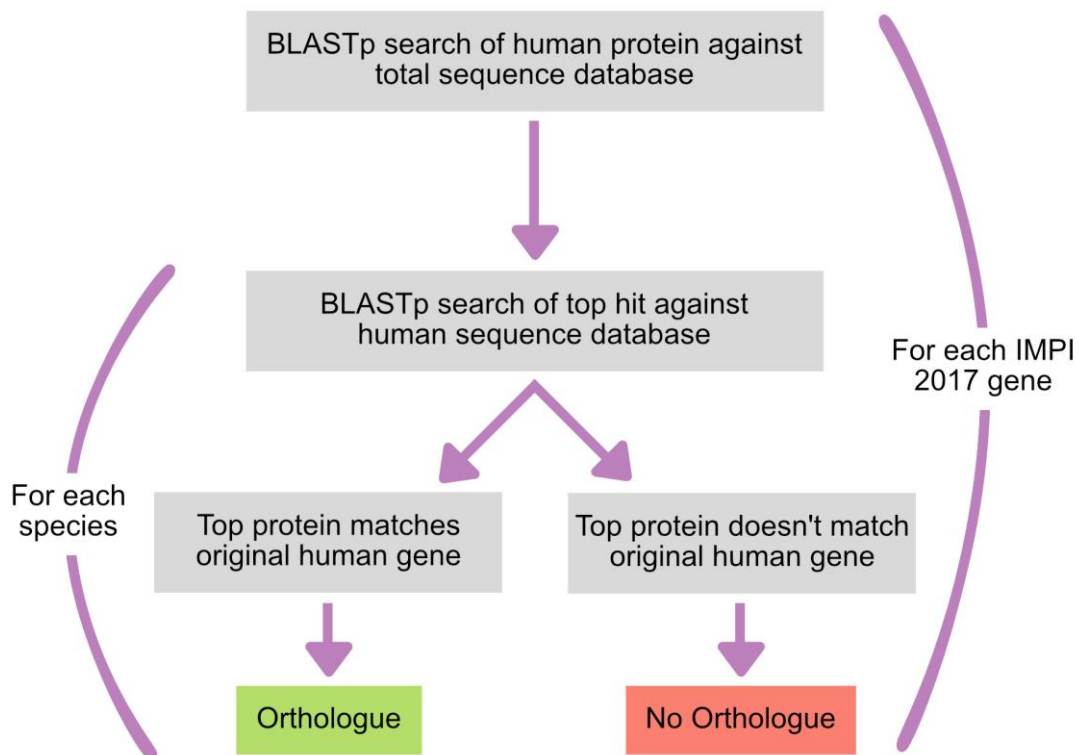


Figure 2.3 Summary of the reciprocal best hit process for the prediction of potential orthologues of IMPI 2017 human proteins.

Identifying and utilising protein domain structure

Protein domains were identified using the NCBI Conserved Domain Database (CDD) (Marchler-Bauer *et al.* 2017). The CDD was used to assign predicted domains to the human protein sequence for each gene and then for all predicted orthologous sequences for each gene, with an E-value threshold of 0.01. Predicted orthologous sequences with domains matching the human sequence were retained as orthologues for further analysis.

Sequences with non-matching domain structures were investigated further. For these sequences, I manually assessed the domain predictions. Sequences without identical matches to the domains from the human protein were kept if they fit at least one of the following criteria:

- The human protein was a known fusion protein and the predicted orthologous sequence matched the domain structure for one half of the fusion protein, or vice versa,
- The human protein and predicted orthologous protein were assigned domains from differing databases included in the total CDD database which are functionally equivalent,
- At least one from the human protein and the predicted orthologous protein were assigned a domain that is specific to its phylogenetic group, but the domains of the two proteins are functionally equivalent,
- The human domain prediction was weak ($E\text{-value} \geq 1 \times 10^{-5}$) and there was no domain prediction for the predicted orthologous sequence,
- The orthologous sequence domain predictions were weak ($E\text{-value} \geq 1 \times 10^{-5}$) and there was no domain prediction for the human protein sequence.

Improving consistency of paralogue assignment

Genes were grouped together based on the results of the reciprocal best hit analysis. For a gene (Gene 1), if the reciprocal BLAST search for any species predicted a non-orthologous sequence, the human gene returned as the top hit (Gene 2) was noted. Gene 1 and Gene 2 were grouped together if the human proteins encoded by these genes contained the same predicted protein domains. If genes from outside IMPI 2017 were identified in these groups,

a reciprocal best hit analysis was carried out for these genes, using the methodology described above.

The phylogenetic patterns of the genes within each of these groups were then manually assessed. Predicted orthologous sequences were assigned to different members of the gene groups as supported by manual assessment, using the support of the literature where available. Where no strong literature evidence was available, orthologues from more distant species were moved to the gene member of the group with the highest number of assigned orthologues, to improve the consistency of assignment for paralogous groups.

Phylogenetic tree

The peroxiredoxin phylogenetic tree was calculated using a variety of sequences from across the tree of life which were predicted orthologues for any of *PRDX1-4*. These sequences were aligned using MUSCLE with default settings (Edgar 2004a, 2004b) and alignments were manually inspected and improved in Jalview 2.0 (Waterhouse *et al.* 2009). A phylogenetic tree was calculated using PhyML 3.0, using the SPR (Subtree Pruning and Regrafting) method and 100 bootstrap replicates (Guindon *et al.* 2010). The tree was visualised in Interactive Tree Of Life (Letunic & Bork 2016).

Gene enrichment

KEGG (Kyoto Encyclopaedia of Genes and Genomes) pathway annotation and enrichment was carried out using the DAVID functional annotation tool (Huang *et al.* 2009b), with IMPI 2017 genes as the background. Benjamini corrected *p*-values from the DAVID analysis were returned, with a *p*-value threshold of 0.05 set for significance.

Assignment of potential gene ancestry

Each gene was assigned to one of six groups, depending on the identification of orthologues in different sets of species. The criteria for each group were implemented in order, as follows:

- *α-proteobacteria* – any gene with at least one orthologue identified in the three studied *α*-proteobacteria (*Agrobacterium tumefaciens* str. C58, *Rhizobium etli* CFN 42, *Rickettsia prowazekii* str. Madrid E),
- *Other bacteria and archaea* – any gene with at least one orthologue in the bacterial species and at least one orthologue in the archaeal species,
- *Other bacteria (not archaea)* – any gene with at least one orthologue in any species of bacteria but not in the *α*-proteobacteria or archaea,
- *Archaea* – any gene with at least one orthologue in the archaeal species, but not in any of the bacterial species,
- *Holozoa* – any gene with orthologues only in the studied Holozoa (animals and the closely related, single-celled ichthyosporea and choanoflagellates),
- *Eukaryota* – the remaining genes, which have no orthologues in any studied prokaryotic species, but have orthologues in eukaryotes outside of the Holozoa.

Results & Discussion

Building an orthology dataset

The aim of the work described in this chapter was to produce a dataset of orthologues of the human mitochondrial proteome, across a range of eukaryotes and prokaryotes. This dataset could then be utilised to investigate the history and function of the human mitochondrial proteome. Therefore, I used reciprocal best hit analysis with BLASTp searches to identify orthologous sequences of the human mitochondrial proteome, across 359 species, including 97 holozoans (animals and their closest single-celled relatives), 134 other eukaryotic species from a variety of taxa, 37 archaea and 91 bacteria (species listed in *Appendix I – Table 1*). The mitochondrial proteome was defined as the canonical proteins encoded by the genes in the human version of IMPI 2017 (www.mrc-mbu.cam.ac.uk/impi), which includes 1,550 individual genes.

The accurate identification of orthologues using BLASTp is dependent on several factors. One of the most varied across the literature is the E-value (Expect value) threshold set for positive results. The E-value is a probabilistic measure of the number of similarly scored hits which would be expected by chance, which is dependent on the size of the database being searched. A lower E-value suggests that a similarly scored result would be less likely to be observed by chance. Increasing the E-value threshold, therefore, increases the number of positive hits from a search, particularly for those sequences which are more distantly related (i.e. increases the sensitivity of the analysis), but also runs the risk of increasing the number of false positives (i.e. decreases the specificity of the analysis).

Therefore, I looked at the effect of changing the E-value of the two different BLASTp searches used in the reciprocal best hit process on the number of predicted orthologues, using E-values ranging between a strict threshold of 1×10^{-30} and a more liberal threshold of 10 (*Table 2.2*). The most liberal tested criteria (E-value of 10 for both BLASTp searches) predicted 36.1% more orthologues than the strictest tested criteria (E-value of 1×10^{-30} for both BLASTp searches). Change in either E-value produced large changes in the number of predicted orthologues.

Table 2.2 Number of predicted orthologues of human proteins encoded by IMPI 2017 genes across 359 species, varying the E-value of the first and reciprocal BLASTp searches.

First BLASTp E-value	Reciprocal BLASTp E-value					
	1×10^{-30}	1×10^{-20}	1×10^{-10}	1×10^{-5}	1	10
1×10^{-30}	149,741	150,131	150,134	150,134	150,136	150,136
1×10^{-20}	152,454	164,421	164,845	164,846	164,849	164,849
1×10^{-10}	152,529	167,140	180,878	181,190	181,213	181,213
1×10^{-5}	152,547	167,180	183,791	180,432	190,688	190,691
1	152,598	167,254	184,035	194,027	201,360	201,415
10	152,604	167,262	184,051	194,101	203,642	203,779

When both E-value thresholds are raised to 10, the increase in the number of predicted orthologues levels off, with only hundreds of additional predictions made compared to the previous tested E-value (1). This suggests that the reciprocal best hit method, using the human sequence only as bait, is reaching its limits in identifying potential additional orthologues. Raising the E-value higher than this is, therefore, not likely to contribute a significant amount – orthologues being missed by the method at this threshold are likely to be missed due to the limitations of the method.

The effect of the E-value threshold was particularly important to consider when attempting to identify orthologous sequences from species which share a more distant common ancestor with humans, such as the prokaryotes, as sequences could be assumed to have diverged more over time. If increasing the E-value threshold increased the number of potential prokaryotic predicted orthologues in particular, then this would justify the use of a higher E-value threshold. To explore how changing the E-values might affect the identification of orthologues in different phylogenetic groups of species, I visualised the number of predicted orthologues in three different phylogenetic groups: the Holozoa (animals and the single-celled ichthyosporea and choanoflagellates), other eukaryotes and prokaryotes (*Figure 2.4*), as the E-value thresholds were varied.

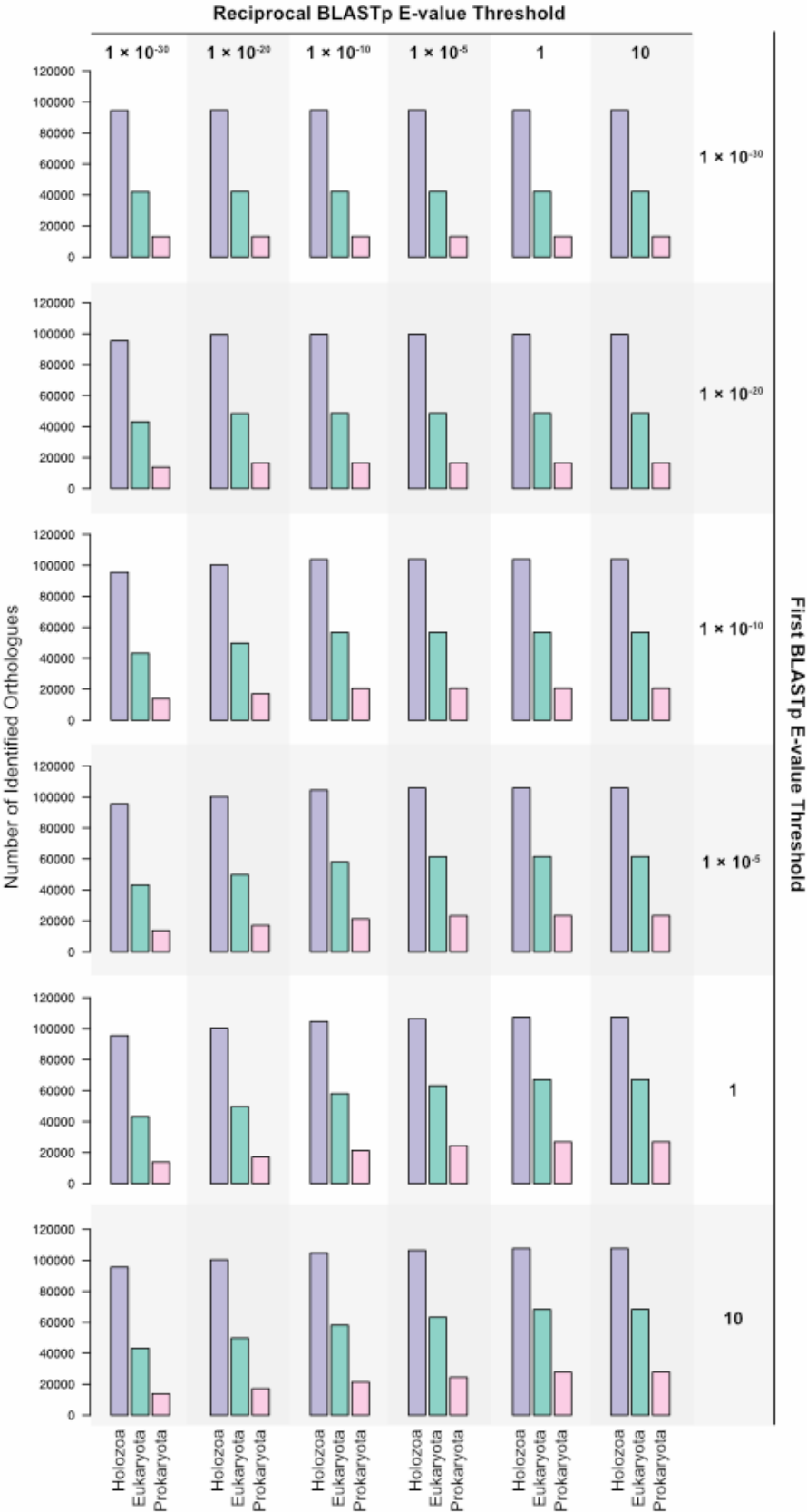


Figure 2.4 Number of predicted orthologues by phylogenetic group, varying the E-value of the BLASTp searches. As the E-value threshold is increased for either BLASTp search, the number of predicted orthologues increases. This is particularly true of the prokaryotic predicted orthologues.

The number of predicted orthologues in eukaryotes increased when increasing the E-value thresholds of both of the first BLASTp search and the reciprocal BLASTp search, but only by a small proportion. The largest proportional change was in the number of predicted prokaryotic orthologues, with an increase in predicted orthologues of 110% when the most liberal E-values were tested (10), compared to the most stringent E-values (1×10^{-30}).

From previous knowledge of orthologue identification, I knew that some orthologous prokaryotic sequences were identified only using high E-value thresholds when searching using the human sequence as bait. For example, the *MT-ATP6* orthologue in *Bacillus subtilis* (P37813.1) has been confirmed experimentally (Santana *et al.* 1994), but is only identified with an E-value of 9.9 for the first BLAST search and 0.015 for the reciprocal BLAST search. From this knowledge and the analysis of E-value thresholds which showed a large increase in prokaryotic predicted orthologues at the most liberal E-values, I decided to use an E-value threshold of 10 (the highest tested threshold) for both the BLASTp searches in the reciprocal best hit process. This is likely to produce a larger number of false positives (orthologues predicted where there is no true orthologous sequence) compared to a more stringent E-value, but minimise the number of false negatives (orthologues not predicted where there is a true orthologous sequence). This is important as the reciprocal best hit process only has average sensitivity amongst orthology detection processes (Chen *et al.* 2007; Dalquen & Dessimoz 2013; Altenhoff *et al.* 2016).

Using protein domains to improve orthologue predictions

As I used liberal E-values for the BLASTp searches to improve the potential sensitivity of the reciprocal best hit analysis, it was possible that the specificity of orthologue prediction may suffer (i.e. that there would be an increased number of false positive orthologue predictions). To attempt to counteract this, I used the idea that proteins with similar domain structures should be stronger candidate orthologues (Li *et al.* 2003). For each gene, predicted orthologous sequences with domain predictions matching that of the human protein sequence for that gene were kept as potential orthologous sequences, using the NCBI's CDD resource to assign protein domains (Marchler-Bauer *et al.* 2017).

Before discarding the remaining sequences, which did not exactly match the domain structure of the human protein, I manually assessed the results for each gene. I used five additional

criteria to assess the potential domain orthology of each sequence – if a predicted sequence fulfilled any one of these criteria, they were kept as potential orthologues and taken forward to the next step. An explanation of each of the criteria with examples is provided below.

The first criterion considered was if the human protein was a bifunctional predicted fusion protein and the predicted orthologue matched the domain structure of one half of the human protein, or vice versa. Fusion proteins usually include two proteins with functions involved in the same pathway (Enright *et al.* 1999), and so I made the assumption that detection of one part of a fusion protein was enough to assume the presence of both parts, as reciprocal best hit analysis would only ever allow the identification of one protein as the top hit. For example, *PAPSSI* encodes a bifunctional protein involved in sulphate activation. The human protein contains two domains corresponding to the two functions of this protein – an ATP-sulfurylase domain (cd00517) and an adenylylsulphate kinase domain (pfam01583). However, the top protein hit for *Saccharomyces cerevisiae* S288c (O43252.2) only encoded an adenylylsulphate kinase domain. Therefore, O43252.2 was assigned as an orthologue of *PAPSSI*, despite encoding only half the domains of the human protein, as it encoded one half of a known fusion protein. The presence of a second protein encoding the ATP-sulfurylase domain was assumed and, indeed, the second hit in a BLASTp search using the human *PAPSSI* sequence is a *S. cerevisiae* protein with an ATP-sulfurylase domain (NP_012543.3).

The second two criteria were similar and considered the differences between domains identified from the different protein domain sources included in CDD. There are some domains in the CDD which are equivalent in function but have different names – either due to coming from different data sources included in the CDD, or there being phylogenetically specific domains within the CDD which share similar function. Proteins assigned functionally equivalent domains were taken forward as predicted orthologues. One example is the identification of an orthologue of the human gene *ATIC* (involved in purine biosynthesis) in *Escherichia coli*. Using CDD search, the human protein contains two domains corresponding to the two functions of this protein – a 5-aminoimidazole-4-carboxamide ribonucleotide transformylase domain (PRK07106) and an inosine monophosphate cyclohydrolase domain (cd01421). However, the *E. coli* protein was predicted with only a single domain (PRK00881: bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/inosine monophosphate cyclohydrolase), which is the functional equivalent of the two domains of the human protein. Therefore, while these two proteins do not produce exactly matching protein domain structures using CDD, they should be, and were, considered orthologous.

The last two criteria were also similar – where one of the protein sequences had a weak domain prediction ($E\text{-value} \geq 1 \times 10^{-5}$) and the potential orthologue had no predicted domain, these proteins were still considered orthologous. This assumed that small variation in sequence may have influenced these domain predictions to cross the E-value threshold either way. For example, the human protein sequence for the gene *C15orf62* has no identified domains using a CDD search at the E-value threshold of 0.001. However, the predicted orthologous sequence in mouse (AAI50993.1) has a single predicted domain (pfam14957), but with an E-value of only 2.6×10^{-4} . Therefore, this protein was assigned as an orthologue of *C15orf62*, despite the predicted protein domains not matching.

In summary, protein domain structure was used to attempt to improve the rate of false positives in the orthology dataset, under the assumption that proteins with similar domain structures are stronger candidate orthologues (Li *et al.* 2003). A set of additional common-sense criteria based on the domain structure were also implemented to reduce the number of true positives excluded from further analysis.

Manually improving paralogue assignment consistency

Despite the use of reciprocal best hit to improve the separation of orthologues and paralogues, it was clear from the presence and absence patterns of predicted orthologues of closely related genes that the method struggled with the consistent assignment of orthology within some closely related human gene groups. This was particularly obvious for sequences from species who shared a common ancestor less recently with humans, whilst the method did reasonably well at separating predicted paralogues in more closely related species (Dalquen & Dessimoz 2013). Therefore, to improve the consistency of assignment of orthologues within these gene groups, I used a manual curation process. This is particularly important for the potential use of this dataset in phylogenetic profiling, which relies on the patterns of presence and absence of orthologues across different species – random scattering patterns of orthologues within gene families will reduce the accuracy of this process.

To address this issue, I considered groups of related genes. These groups were based on human genes which returned reciprocal best hits to each other – i.e. genes were grouped together if the reciprocal best hit for an original BLASTp search for Gene 1 instead returned

Gene 2 – and had identical predicted protein domain structures. This included genes not present in IMPI 2017 that were identified as a best hit. I then manually considered the phylogenetic patterns of presence and absence of each of the genes in each group, across all species. Orthologue assignment within these groups was altered where deemed appropriate to improve the consistency of orthologue identification both within and between phylogenetic groups. These changes were made so that, where the phylogenetic pattern of presence and absence of different genes within the gene group was inconsistent across species, a single human gene in the group was chosen, and orthologues from more distantly related species were assigned to this gene. This change worked under the assumption that the gene group developed from a single ancestral gene, which should, therefore, be represented by a single member of the gene group. Where possible, these decisions were made based on information about gene family development from the literature.

One example of deconvoluting the orthologues in a gene family is provided by the three paralogous human genes encoding subunit *c* of the ATP synthase (*ATP5G1*, *ATP5G2*, and *ATP5G3*), which were formed by gene duplication (Abbasi 2010). *Figure 2.5(a)* shows the original gene assignments of the orthologous sequences from all tested species with at least one predicted orthologue. Despite the erratic phylogenetic patterns of presence and absence of orthologues across these three genes, the highest number of orthologues are assigned to *ATP5G2*. This is consistent with the literature evidence from Abbasi (2010) of *ATP5G2* branching earliest from the ancestral gene sequence, before the duplication and separation of the *ATP5G1/3* ancestral sequence into the two genes now observed in humans (Abbasi 2010). Therefore, orthologues in species where only one *ATP5G* orthologue was predicted were assigned to *ATP5G2*, creating a consistent phylogenetic pattern across species. The improvement in consistency of gene assignment is visible in *Figure 2.5(b)*, which shows the manually curated assignment of predicted orthologues of the *ATP5G* genes across species.

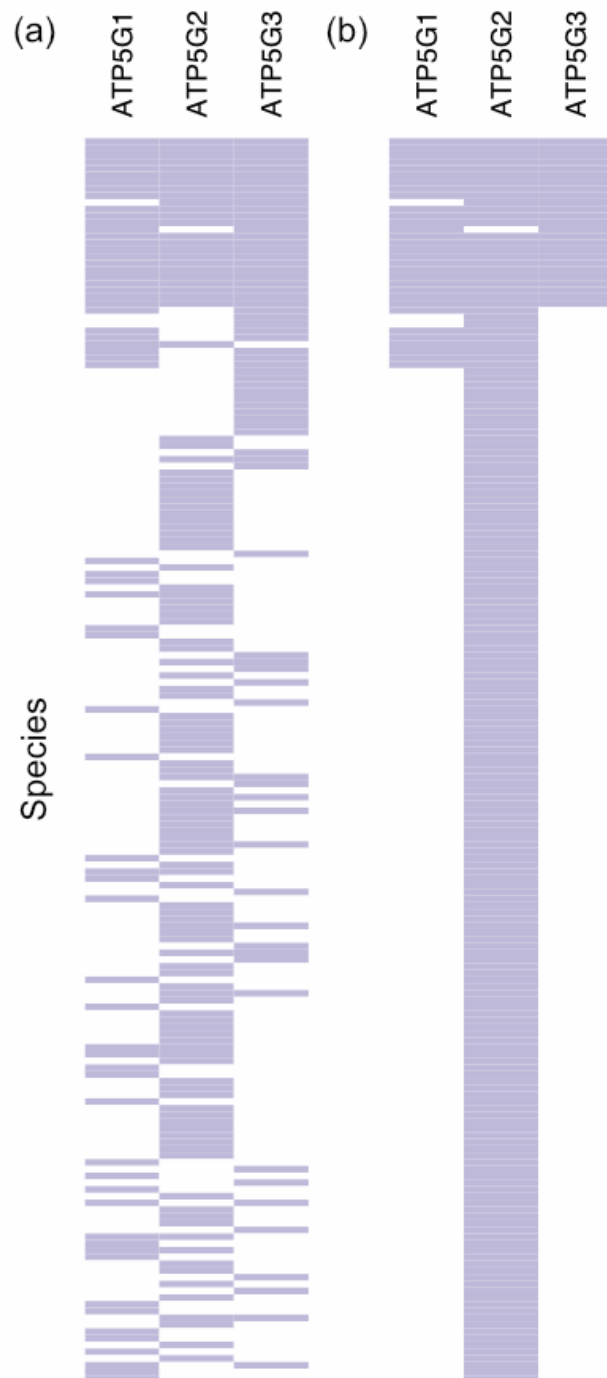


Figure 2.5 Phylogenetic assignment of ATP5G predicted orthologues across all species with at least one predicted orthologue. Each block represents a predicted orthologue of the labelled gene for one investigated species. (a) Original assignments after reciprocal best hit analysis and protein domain checks. (b) Corrected assignments after manual improvement, showing the greater consistency across species.

Another, more complex, example is provided by the 2-Cys class of peroxiredoxins, which includes the human genes *PRDX1*, *PRDX2*, *PRDX3* and *PRDX4* – all included in the IMPI 2017 mitochondrial gene inventory. *Figure 2.6* shows a bootstrapped phylogenetic tree of some of the most diverse predicted orthologous sequences of proteins encoded by these four genes, coloured by their original gene assignment from the reciprocal best hit analysis. It is clear from the tree that the reciprocal best hit method found it difficult to consistently assign sequences as orthologues of genes, particularly for species which are more distantly related to humans. Potential *PRDX4* orthologues from holozoans grouped reasonably well and so were separated from the other sequences, as did *PRDX3* potential orthologues from the Holozoa (*Figure 2.6*). The tree also suggests that *PRDX1* separated off from *PRDX2*, with sequences representing orthologues of these two genes clustered together.

In this case, there was little literature evidence investigating the history of this group of genes, particularly in relation to the human genes which are most of interest in this analysis. Without this knowledge and with no strong evidence from the tree, the more distant sequences were assigned to *PRDX2* to maintain consistency across species, as more sequences had hit back to *PRDX2* than any of the other genes. This assignment was made, as before, under the assumption that the more distantly related sequences (from species beyond the Holozoa) developed from one ancestral sequence. Predicted orthologous sequences from holozoan species were kept at their original assigned gene.

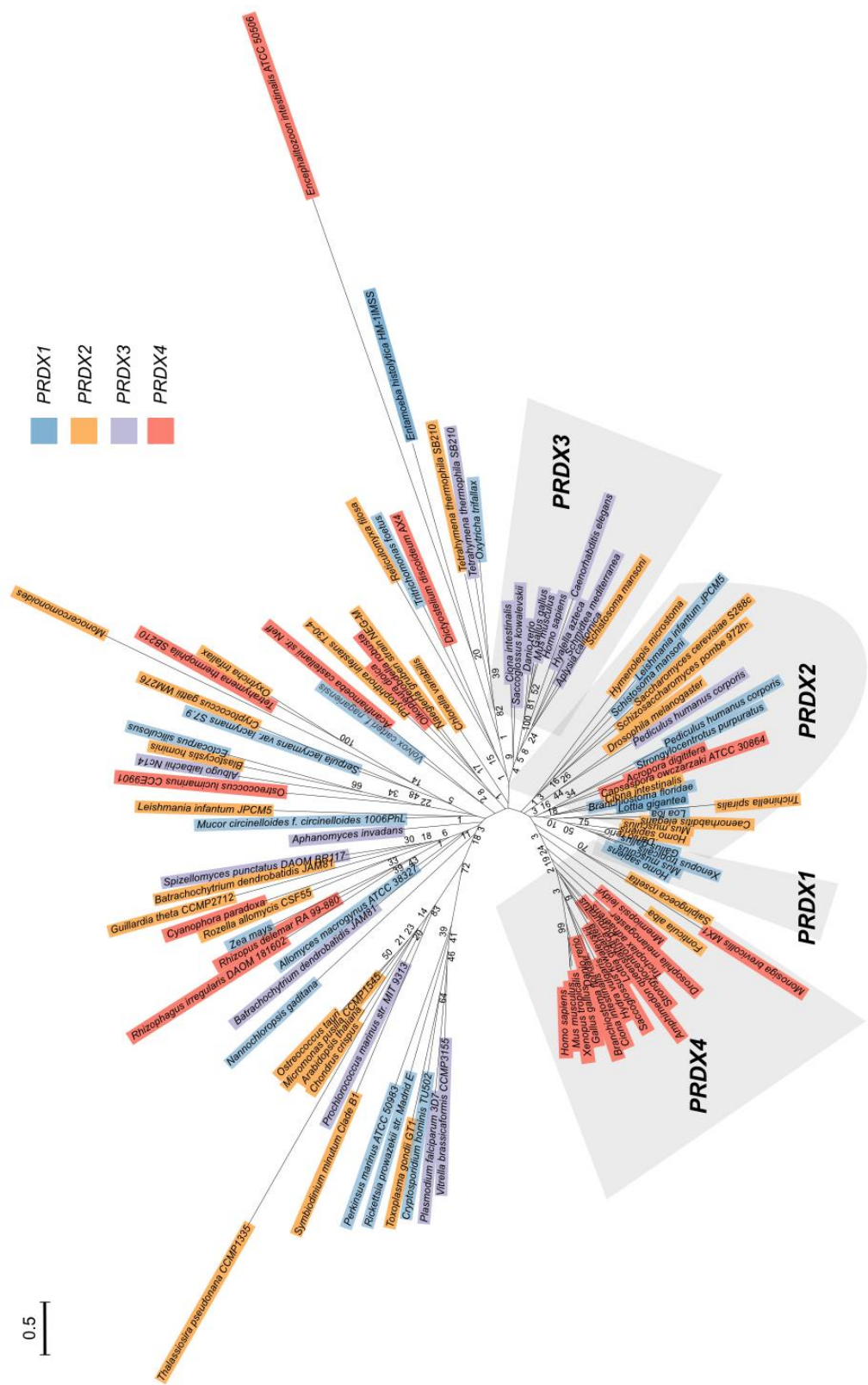


Figure 2.6 Phylogenetic tree of a variety of predicted orthologous sequences of human PRDX1-4, coloured by their original gene assignment after reciprocal best hit analysis (as described in the key). Sequences from species more distantly related to humans are returned as orthologues to any of PRDX1-4, despite there likely being only one ancestral sequence. Branch length scale indicates number of substitutions per site.

This example shows the difficulty with assigning orthologous sequences predicted from distantly related species within groups of related genes using the reciprocal best hit methodology. Literature knowledge, manual assessment and simple phylogenetic trees can inform the assignment in some cases. but the history of most of these gene groups, in relation to humans, remains obscure, and would justify a full and further analysis to understand the homology relationships of these genes across the tree of life.

This protocol assumes a low rate of horizontal gene transfer. This assumption is more likely accurate within the eukaryotes than within prokaryotes due to lower rates of horizontal gene transfer, although horizontal gene transfer is known within the eukaryotes (Keeling & Palmer 2008). It also does not consider paralogous sequences in species apart from those in humans, grouping all these together under the representation of one sequence from each species. This itself is a general limitation of reciprocal best hit analysis, contributing to the method's lower sensitivity levels (Dalquen & Dessimoz 2013). However, this weakness is not relevant to the success of this particular analysis, which is concerned with how the human mitochondrial proteome developed over time – not in how families of genes developed over time in different species.

In this analysis, consistency across species was prioritised, with a single human gene within each gene group chosen to assign orthologous sequences from more distantly related species, where there was less knowledge of the gene history. This assumption should be considered when using this dataset for further analysis.

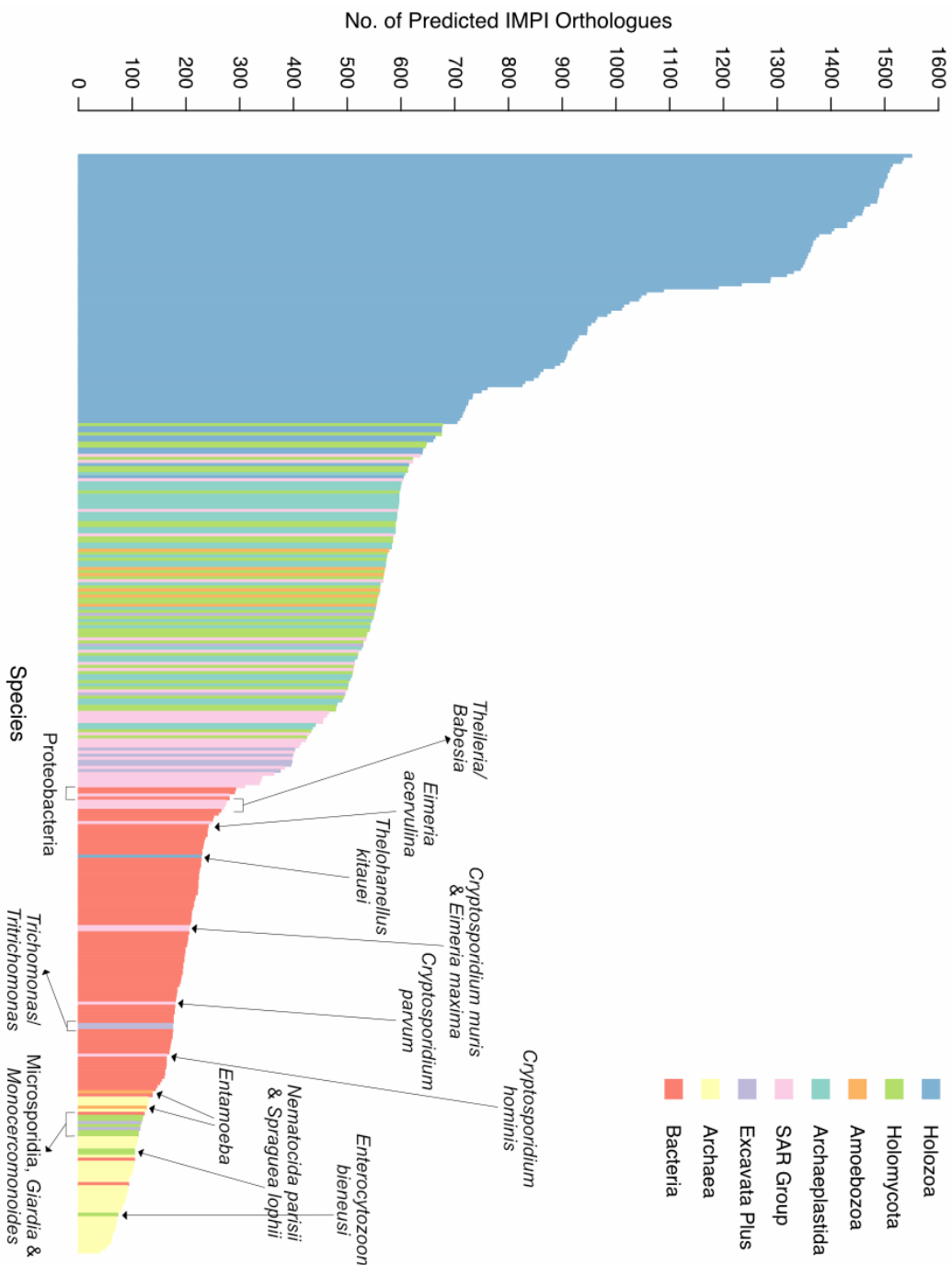
Final orthology dataset

The final orthology dataset contains 190,097 predicted orthologues of the 1,550 IMPI 2017 genes, across 359 species of eukaryota, archaea and bacteria. This is a net loss of 13,682 predicted orthologous sequences from the original reciprocal best hit analysis, after protein domain analysis and manual curation of orthologue assignment within gene groups. Over half (56.1%) of the total number of predicted orthologues are from the 97 studied holozoans, with 32.7% of the predicted orthologues from the 132 other eukaryotic species. Only 1.7% of all the predicted orthologues are from the 37 archaea and 9.5% of all the predicted orthologues are from the 92 bacteria.

The species with the highest number of predicted orthologous sequences were other primates (see *Appendix I – Table 1*). Orthologues of 98.8% of human IMPI 2017 genes were predicted in *Pan troglodytes* – the most closely related species to humans studied in this analysis. The studied species with fewest predicted orthologues were archaea (*Figure 2.7*), with only 2.6% of IMPI 2017 genes identified in the archaeal species *Nanoarchaeum equitans* Kin4-M. The prokaryotic species with the highest number of predicted orthologues was *Rhizobium etli* – one of the three studied α -proteobacteria. This is consistent with the proposed origin of the mitochondria as an α -proteobacterium in endosymbiosis with a host cell (Fitzpatrick *et al.* 2006).

Twenty-four of the studied eukaryotic species had fewer identified orthologues than *Rhizobium etli* (labelled on *Figure 2.7*). Twenty-three of these species are known to contain derivative remnants of full mitochondria. The eukaryote with the fewest predicted IMPI 2017 orthologues is *Enterocytozoon bieneusi* H348, a member of the Microsporidia – a division of parasitic fungi. These species contain highly derived forms of mitochondria called mitosomes (Vávra 2005). *E. bieneusi* is known to have a highly reduced genome even among the Microsporidia (Akiyoshi *et al.* 2009), which is reflected in the number of predicted IMPI 2017 orthologues.

Figure 2.7 Number of IMPI 2017 predicted orthologues for each of the 359 studied species, coloured by the phylogenetic taxon of each species. Eukaryotic species with lower numbers of predicted IMPI orthologues than at least one prokaryote are labelled.



The one exception, which is not known to have derivative mitochondria, is a holozoan species – *Thelohanellus kitauei* – which has only 228 predicted orthologues of the 1,550 studied IMPI 2017 genes (14.7%). *T. kitauei* is a member of the Myxozoa (Yang *et al.* 2014) – a poorly studied taxon consisting of obligate parasites. The Myxozoa have recently been assigned to the taxon Cnidaria, which includes jellyfish, sea anemones, corals and box jellies (Chang *et al.* 2015). Considering the paucity of mitochondrial proteome orthologues identified in *T. kitauei*, it could be interesting to investigate the structure and function of mitochondria in this species and/or in other Myxozoa.

Evolutionary history of the human mitochondrial proteome

Mitochondria have a unique evolutionary history in eukaryotes, as they are thought to be the result of an endosymbiotic relationship of an α -proteobacterium with another cell (Sagan 1967). The subsequent development of mitochondria has led to a proteome with a wide-ranging evolutionary history, ranging from genes predicted to have been inherited from the original α -proteobacterial endosymbiont, to genes which are novel to eukaryotes (Szklarczyk & Huynen 2010). Most previous work analysing the origins of the mitochondrial proteome has been based on the yeast *Saccharomyces cerevisiae* (Kurland & Andersson 2000; Karlberg *et al.* 2000; Gray 2015). However, only 30.8% of IMPI 2017 genes have predicted orthologues in *S. cerevisiae*. One simple analysis of the potential origins of the human mitochondrial proteome has been carried out (Pagliarini *et al.* 2008). This analysis was performed using only single BLASTp searches, as well as examining only a few eukaryotic genomes. It is, therefore, interesting to consider the possible origins of genes encoding the human mitochondrial proteome, using the newly derived set of orthologues from a wide range of species.

There is a wide variety in the number of predicted orthologues for each of the IMPI 2017 genes (*Figure 2.8*). The 10% of IMPI 2017 genes with the highest numbers of predicted orthologues are significantly enriched for two KEGG pathways at the 0.05 significance level: ‘Ribosome’ ($p = 1.1 \times 10^{-6}$) and ‘Aminoacyl-tRNA biosynthesis’ ($p = 8.3 \times 10^{-3}$). The bottom 10% of genes from IMPI 2017 with the lowest numbers of predicted orthologues are not enriched for any KEGG pathway.

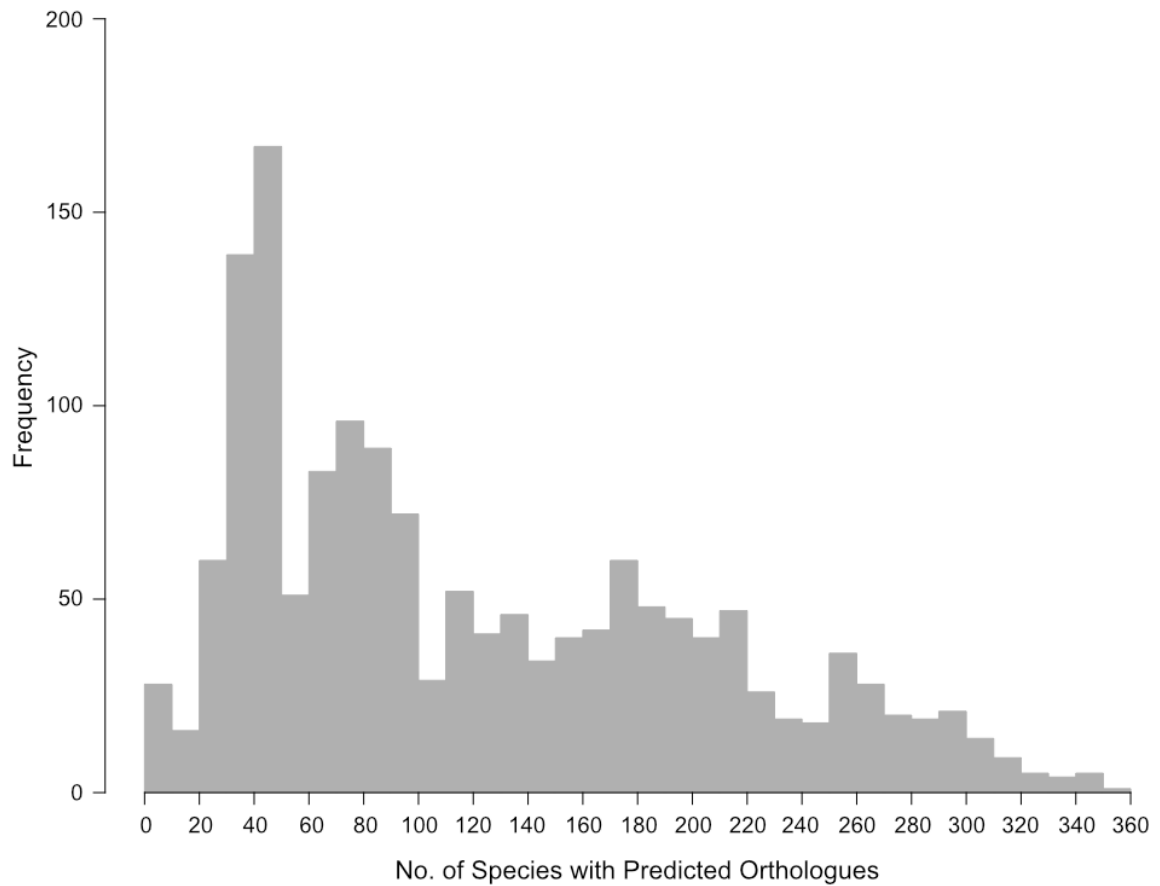


Figure 2.8 Histogram of the number of predicted orthologues identified for each of the 1,550 human IMPI 2017 genes, in 359 species.

There are several peaks of gene innovation (*Figure 2.8*):

- the first peak of genes with orthologues predicted in approximately 40 species
- the second peak at around 80 species
- the third, smaller and less distinct peak at around 180 species
- and a fourth and final peak at around 260 species, which tails off to the right.

By looking at the breakdown of the number of orthologues identified in different phylogenetic groups for each gene, it is possible to estimate what these numbers mean in terms of the types of species with predicted orthologues of these genes (*Figure 2.9*).

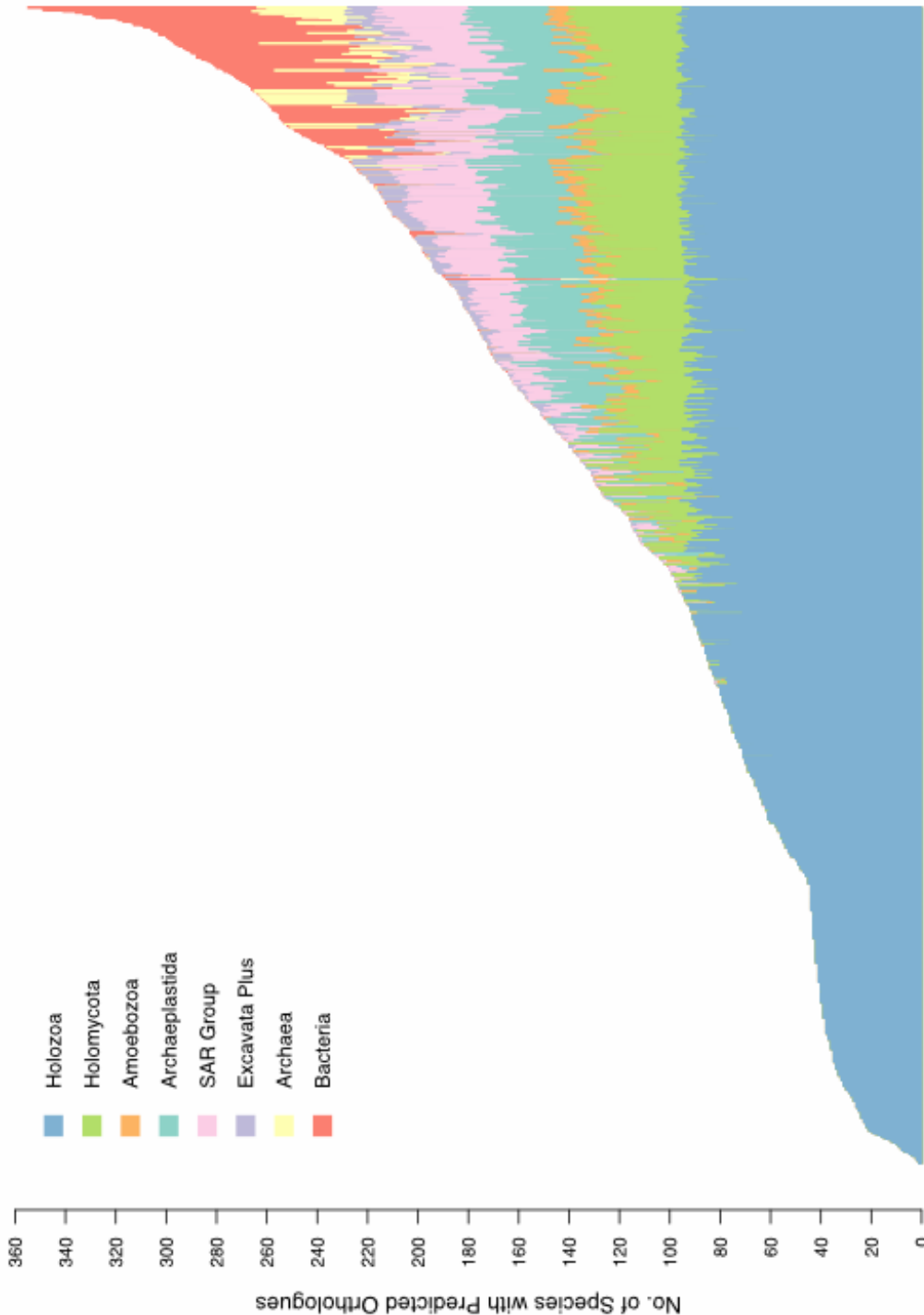


Figure 2.9 Breakdown of the number of orthologues predicted from each of eight phylogenetic taxa, for each of the 1,550 genes in IMPI 2017, ordered by the total number of predicted orthologues. Each bar represents one IMPI 2017 gene. The height of the bar indicates the number of predicted orthologues. The colour of the bar represents the phylogenetic taxon of the predicted orthologues, as described in the key.

The first peak (~40 species) is equivalent to genes which are new innovations in the vertebrates. The second peak (~80 species) is equivalent to genes which are novel to the Holozoa – animals and closely related single-celled organisms. The third peak (~180 species) is equivalent to genes which are eukaryotic innovations. The fourth peak (~260 species) contains only genes that have predicted orthologues in at least some of the prokaryotic species studied. Overall, this suggests that there have been several important points in evolutionary time for the evolution of the human mitochondrial proteome, two of which potentially occurred post the split of Holozoa from fungi (and, therefore, post the split with *S. cerevisiae*, which has been the mitochondrial proteome studied most often).

I decided to breakdown the potential origins of IMPI 2017 genes further by using a simple assignment process, based on the prediction of orthologues in defined groups in a linear assignment process (*Figure 2.10*), starting from assignment to the α -proteobacterial group.

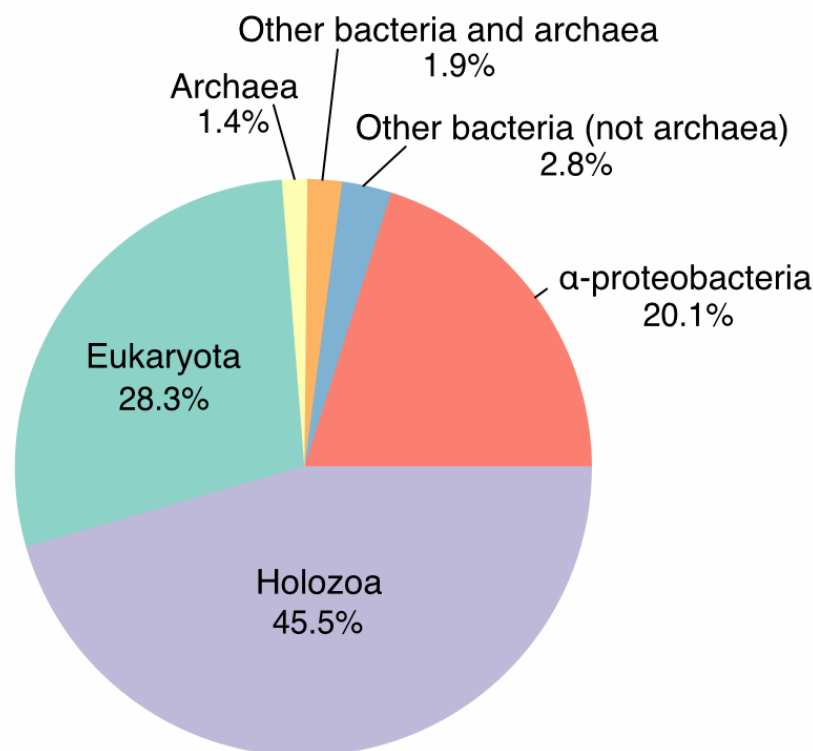


Figure 2.10 Potential ancestry of IMPI 2017 genes. Genes are assigned to a group based on the prediction of orthologues within certain taxa. Any gene with a predicted orthologue in the α -proteobacteria are assigned to the group ‘ α -proteobacteria’. Genes with other prokaryotic predicted orthologues are assigned to the corresponding prokaryotic group. Genes with predicted orthologues only in holozoan species are assigned to the group ‘Holozoa’, with all remaining genes assigned to the group ‘Eukaryota’.

Genes with at least one orthologue identified in one of the three α -proteobacteria investigated were assigned to the α -proteobacterial group, assuming that these genes originated from this group as the probable origin of the mitochondria (Müller & Martin 1999). This is probably a simplification of the true contribution of α -proteobacteria to the mitochondrial proteome for several reasons. Genes may have been transferred to α -proteobacteria after the mitochondrial endosymbiotic event. Genes may, also, have been present in other α -proteobacteria at the time, but not the specific and unknown endosymbiotic ancestor. Lastly, genes may have been lost from the endosymbiont and regained from another source later in evolutionary history. Despite these issues and the inclusion of only three α -proteobacteria in the dataset, 312 IMPI 2017 proteins (20.1%) have at least one identified α -proteobacterial orthologue, with 280 (18.1%) identified in at least two α -proteobacteria. This is consistent with previous analyses using different methods and descriptions of the mitochondrial proteome, which have suggested between 10 and 20% of the eukaryotic mitochondrial proteome is derived from α -proteobacteria (Karlberg *et al.* 2000; Gabaldón & Huynen 2007; Wang & Wu 2014).

Genes with predicted orthologues in other prokaryotes but not α -proteobacteria were then assigned to one of three groups. 29 human genes had predicted orthologues in bacteria and archaea. 43 human genes had predicted orthologues in other bacteria but not archaea. 22 human genes had predicted orthologues in archaea but not in bacteria. In total, 198 genes had predicted orthologues in at least one bacteria but not in archaea. This is probably reflective of the presumed origin of the mitochondria as a bacterial endosymbiont.

Genes with at least one predicted prokaryotic orthologue were enriched for several pathways (Table 2.3). These results suggest the key parts of the mitochondrial proteome inherited from prokaryotic species were the mitochondrial ribosome (known to be at least partially inherited from bacteria (O'Brien 2002)) and the basic metabolic pathways that are localised to the mitochondria in humans, such as the tricarboxylic acid cycle, and amino acid biosynthesis.

Table 2.3 Enriched KEGG pathways for IMPI 2017 genes with predicted prokaryotic orthologues (corrected p-value threshold of 0.05).

KEGG pathway	Count	p-value	Benjamini corrected p-value
Ribosome	54	3.2×10^{-16}	4.8×10^{-14}
Metabolic pathways	156	3.7×10^{-4}	0.03
Biosynthesis of antibiotics	20	8.4×10^{-4}	0.04
Carbon metabolism	34	1.2×10^{-3}	0.04
Biosynthesis of amino acids	21	1.5×10^{-3}	0.04

In total, 1,143 (73.8%) IMPI proteins were only identified in eukaryotes. This is a larger proportion than in previous analyses of mitochondrial proteins of eukaryotic organisms, where between 20 – 54% of identified proteins had a predicted eukaryotic origin (Marcotte *et al.* 2000; Smith *et al.* 2007; Pagliarini *et al.* 2008; Szklarczyk & Huynen 2010). These genes were then divided into those with orthologues only in holozoan species (45.5%), and those with predicted orthologues in at least one of the other studied eukaryotes (28.3%). It is surprising that nearly half of the genes encoding the human mitochondrial proteome are novel to the Holozoa. KEGG pathway enrichment analysis shows that these genes are enriched for a wide range of processes (Table 2.4), but these processes, in general, tend to be involved in signalling, complex diseases and the immune system. An increase in the complexity and number of signalling pathways has previously been noted in animals compared to other related species (Suga *et al.* 2013).

Table 2.4 Enriched KEGG pathways for IMPI 2017 genes with only holozoan orthologues (corrected p-value threshold of 0.05).

KEGG pathway	Count	p-value	Benjamini corrected p-value
Hepatitis B	16	1.4×10^{-5}	3.3×10^{-3}
MicroRNAs in cancer	16	4.7×10^{-5}	5.6×10^{-3}
PI3K-Akt signalling pathway	17	1.4×10^{-4}	0.01
Ras signalling pathway	13	2.4×10^{-4}	0.01
Progesterone-mediated oocyte maturation	11	4.8×10^{-4}	0.02
Gap junction	11	4.8×10^{-4}	0.02
Oestrogen signalling pathway	12	6.0×10^{-4}	0.02
Focal adhesion	14	7.1×10^{-4}	0.02
Proteoglycans in cancer	14	7.1×10^{-4}	0.02
Insulin signalling pathway	14	7.1×10^{-4}	0.02
Prolactin signalling pathway	10	1.2×10^{-3}	0.04
Viral carcinogenesis	16	1.4×10^{-3}	0.04
MAPK signalling pathway	12	1.6×10^{-3}	0.04

In summary, this analysis suggests that the human mitochondrial proteome has been formed over time from a mosaic of different sources, as has been suggested for the yeast mitochondrial proteome (Gray 2015). The mitochondrial proteome has continued developing even within the Holozoa, likely expanding and complicating the functions of mitochondria within the cell, far beyond the traditional view of the mitochondrial energy ‘powerhouse’.

Conclusions

In this chapter, I have described the process of creating a dataset of orthologues of the IMPI 2017 genes predicted to encode the human mitochondrial proteome, across 359 eukaryotes, archaea and bacteria. Reciprocal best hit analysis was used to identify the initial orthologous sequences. This was then improved by utilising the protein domain structure of the sequences. Finally, manual analysis of closely related groups of genes was used to improve the consistency of assignments among these groups. This provides a resource for furthering our understanding of the history of the human mitochondrial proteome, as well for other phylogeny-based analyses to understand the function of particular genes in the mitochondria.

I then used this dataset to study the potential origins of the genes encoding the human mitochondrial proteome. A surprising number of these genes were recent innovations, unique to the Holozoa, suggesting a large increase in complexity of the mitochondria even within animals. This is interesting for an organelle which has traditionally been viewed mainly in terms of basic energy synthesis and metabolism.

Chapter 3

Investigating the mitochondrial respiratory complexes using phylogenetic profiling

Introduction

Assembly of the mitochondrial respiratory chain complexes

The four complexes of the mitochondrial respiratory chain and ATP synthase are each formed from a combination of different protein subunits and cofactors. These protein complexes may assemble to form supercomplexes – groups of these functionally related complexes in the inner mitochondrial membrane (Schägger & Pfeiffer 2000). A group of proteins known as assembly factors help to coordinate the formation of these complex structures (Mckenzie & Ryan 2010; Smith *et al.* 2012).

Complex I (NADH dehydrogenase) is the largest of the four respiratory complexes, consisting of forty-five subunits in humans – fourteen highly evolutionarily conserved core subunits and thirty-one supernumerary subunits (Zhu *et al.* 2016). These subunits form an L-shaped assembly, with a hydrophobic arm sitting in the inner mitochondrial membrane, and a hydrophilic arm pointing into the mitochondrial matrix (Zhu *et al.* 2016). The complex I structure also includes cofactors – a flavin mononucleotide as well as several iron-sulphur clusters (Zhu *et al.* 2016). There are currently at least fourteen assembly factors known to support the assembly and stability of mammalian complex I (Guerrero-Castillo *et al.* 2017). The size and complexity of the complex I structure and the existence of complex I deficient patients without a current genetic diagnosis even after exome sequencing (Calvo *et al.* 2012) suggests that there may be additional assembly factors still to be identified (Andrews *et al.* 2013). Even the electron transfer flavoprotein, which also moves electrons into the electron transport chain, but is formed from only two subunits (Roberts *et al.* 1996), has two known assembly factors (Małeck *et al.* 2015; Floyd *et al.* 2016).

One potential method to identify genes with functions associated with complex I and the other respiratory complexes is phylogenetic profiling.

Phylogenetic profiling

There are many different ways to predict the function of a protein. Phylogenetic profiling is a technique which uses the presence and absence of genes across species to predict the possible function of an uncharacterised protein. Proteins with related functions tend to have similar patterns of presence and absence within species (Pellegrini *et al.* 1999; Loganantharaj & Atwi 2007). While this technique has been widely and successfully applied in prokaryotic species, usage in eukaryotic species is limited by the decreased variation in the phylogenetic profiles of genes and the larger number of paralogous gene families (Snitkin *et al.* 2006). Successful use of this technique with eukaryotic species therefore relies on good orthology prediction, and variation in the presence and absence of the complex or pathway of interest between related species.

A paper from the Mootha group used phylogenetic profiling to identify potential complex I assembly factors, with two of the highest rated predictions being functionally confirmed (Pagliarini *et al.* 2008). Whilst the orthology detection technique used in this study was simple (a single BLASTp search), there was a strong phylogenetic pattern for the subunits of complex I. This includes several related yeast species with a loss of this complex, which was also noted in additional parasitic eukaryotes. The expansive orthology dataset described in *Chapter 2* provided an opportunity to repeat this type of analysis on a wider range of mitochondrial proteins and species, and to look beyond complex I at other respiratory complexes, the electron transfer flavoprotein and ATP synthase.

Chapter summary

In this chapter, I describe the use of the mitochondrial proteome orthology dataset and the technique of phylogenetic profiling to investigate the mitochondrial proteome for potential assembly factors of the respiratory complexes, particularly complex I and the electron transfer flavoprotein.

Methods

Phylogenetic profiling of complex I

A binary matrix of presence and absence of predicted orthologues of the 1,550 IMPI 2017 genes across 359 species was created from the orthology dataset described in *Chapter 2*. This was used to facilitate the identification of genes fitting different phylogenetic profiles for the respiratory chain complexes, electron transfer flavoprotein and ATP synthase.

For complex I, COPP Tier 1 (Complex I Phylogenetic Profile Tier 1) genes were identified using a slightly modified version of the phylogenetic pattern defined in Pagliarini *et al.* 2008. The pattern for the identification of COPP1 genes was as follows:

- Absent in: *Schizosaccharomyces pombe* 972h-, *Candida glabrata* CBS 138, *Saccharomyces cerevisiae* S288c, *Cryptosporidium hominis* TU502, *Cryptosporidium parvum* Iowa II, *Plasmodium falciparum* 3D7, *Theileria annulata* strain Ankara, *Theileria parva*, *Giardia lamblia* ATCC 50803 and *Encephalitozoon cuniculi* GB-M1. (Does not include *Ashbya gossypii* which was in the original paper, as this species was not included in the orthology dataset).
- Present in a bacterial genome.
- Present in at least one of the plant-like species: *Arabidopsis thaliana*, *Oryza sativa* Japonica group, *Dictyostelium discoideum* AX4 and *Cyanidioschyzon merolae* strain 10D.
- Present in at least two of the following yeast species: *Yarrowia lipolytica* CLIB122, *Candida albicans* SC5314, *Scheffersomyces stipitis* CBS 6054 and *Debaryomyces hansenii* CBS767.

The newly designated COPP Tier 1B genes were identified using the same criteria, except removing the requirement for identification of at least one bacterial orthologue.

Further complex I deficient species were identified by looking at the presence and absence of the core complex I subunits (*MT-ND1*, *MT-ND2*, *MT-ND3*, *MT-ND4*, *MT-ND4L*, *MT-ND5*, *MT-ND6*, *NDUFS1*, *NDUFS2*, *NDUFS3*, *NDUFS7*, *NDUFS8*, *NDUFV1* and *NDUFV2*) for each species in the orthology dataset. Species missing all or most of the core subunits were designated as complex I deficient, confirmed by the literature where possible.

The full complex I phylogenetic profile, therefore, also included the additional criteria of absence from the following complex I deficient species: *Nematocida parisii* ERTm3, *Edhazardia aedis* USNM 41457, *Vavraia culicis*, *Trachipleistophora hominis*, *Spraguea lophii* 42_110, *Encephalitozoon intestinalis* ATCC 50506, *Enterocytozoon bieneusi* H348, *Rozella allomyces* CSF55, *Entamoeba dispar* SAW760, *Entamoeba histolytica* HM-1:IMSS, *Vitrella brassicaformis* CCMP3155, *Symbiodinium minutum* Clade B1, *Babesia bovis* T2Bo, *Eimeria tenella*, *Eimeria maxima*, *Eimeria acervulina*, *Cryptosporidium muris* RN66, *Theileria orientalis* strain Shintoku, *Toxoplasma gondii* GT1, *Neospora caninum* Liverpool, *Plasmodium chabaudi chabaudi*, *Trichomonas vaginalis* G3, *Tritrichomonas foetus* and *Monocercomonoides*.

Electron transfer flavoprotein phylogenetic profiling

Eukaryotic species predicted to be missing electron transfer flavoprotein were identified as those without predicted orthologues of the two genes encoding the subunits of the electron transfer flavoprotein (*ETFA* and *ETFB*) and the essential associated dehydrogenase (*ETFDH*).

These species were: *Schistosoma mansoni*, *Schistosoma haematobium*, *Hymenolepis microstoma*, *Echinococcus granulosus*, *Echinococcus multilocularis*, *Thelohanellus kitauei*, *Pneumocystis jirovecii* RU7, *Nematocida parisii* ERTm3, *Edhazardia aedis* USNM 41457, *Vavraia culicis*, *Trachipleistophora hominis*, *Spraguea lophii* 42_110, *Encephalitozoon intestinalis* ATCC 50506, *Encephalitozoon cuniculi* GB-M1, *Enterocytozoon bieneusi* H348, *Entamoeba dispar* SAW760, *Entamoeba histolytica* HM-1:IMSS, *Ostreococcus tauri*, *Ostreococcus lucimarinus* CCE9901, *Micromonas pusilla* CCMP1545, *Babesia bovis* T2Bo, *Eimeria tenella*, *Eimeria maxima*, *Eimeria acervulina*, *Cryptosporidium hominis* TU502, *Cryptosporidium muris* RN66, *Cryptosporidium parvum* Iowa II, *Theileria annulata* strain Ankara, *Theileria parva*, *Theileria orientalis* strain Shintoku, *Plasmodium chabaudi chabaudi*, *Plasmodium falciparum* 3D7, *Trichomonas vaginalis* G3, *Tritrichomonas foetus*, *Giardia lamblia* ATCC 50803 and *Monocercomonoides*.

The phylogenetic profile for identification of potential electron transfer flavoprotein associated genes was then defined as follows:

- Absent in all eukaryotic species predicted to lack electron transfer flavoprotein.
- Present in at least two of the free-living cnidarians: *Nematostella vectensis*, *Acropora digitifera* and *Hydra vulgaris*. This criterion was included to exclude genes that were not identified in any cnidarians, rather than specifically lost in *Thelohanellus kitauei*.
- Present in at least two of the following platyhelminthes: *Schmidtea mediterranea*, *Clonorchis sinensis* and *Opisthorchis viverrini*. This criterion was included to exclude genes that were not identified in any platyhelminthes, rather than specifically lost in blood flukes and tapeworms.

Complexes II-IV and ATP synthase phylogenetic profiling

Eukaryotic species missing each of the complexes II-IV and ATP synthase were identified as those missing all subunits of each of the considered complexes. The subunits of the complexes considered were identified from the HGNC (HUGO Gene Nomenclature Committee) gene families (Gray *et al.* 2016) as listed in *Table 3.1*. Known assembly factors were also included and were identified from the literature.

Table 3.1 HGNC gene families for the respiratory complexes II-IV and ATP synthase. Loss of the genes within these families was used to inform the phylogenetic profiles for these complexes.

Complex	HGNC gene family	No. of subunits
Complex II	Mitochondrial complex II: succinate dehydrogenase subunits	4
Complex III	Mitochondrial complex III: ubiquinol-cytochrome <i>c</i> reductase complex subunits	10
Complex IV	Mitochondrial complex IV: cytochrome <i>c</i> oxidase subunits	19
ATP synthase	Mitochondrial complex V: ATP synthase subunits	20

Results & Discussion

Phylogenetic profiling of complex I (NADH dehydrogenase)

Phylogenetic profiling uses the phylogenetic pattern of genes in a presence and absence matrix to associate genes with certain pathways or functions (Loganathanaraj & Atwi 2007). Previous studies had used this technique to identify assembly factors of complex I (Ogilvie *et al.* 2005; Pagliarini *et al.* 2008). Given the large spread of species investigated and the increased accuracy of orthologue prediction expected from the orthology dataset described in Chapter 2, I decided to explore the use of phylogenetic profiling in identifying additional respiratory chain complex assembly factors.

Initially, I applied this technique to complex I – the largest of the four respiratory chain complexes and the first complex of the electron transport chain (Vinothkumar *et al.* 2014). I first repeated the analysis from the Mootha group paper which had identified two complex I assembly factors (Pagliarini *et al.* 2008), to compare any changes brought about by using the IMPI 2017 dataset and a different orthology prediction method. This paper identified potential complex I associated genes known as ‘COPP Tier 1’ (Complex I Phylogenetic Profile Tier 1) genes. These genes followed a specific pattern of presence and absence across species, including loss in several species known to lack complex I, presence in at least one bacteria and one plant-like species, and presence in at least two yeast species which are known to contain complex I (Pagliarini *et al.* 2008).

Twenty-one IMPI 2017 genes fit the COPP Tier 1 phylogenetic pattern: thirteen known complex I subunits, three known complex I assembly factors and five other genes (highlighted in green and yellow in Figure 3.1). This included one known complex I assembly factor – *NUBPL* (Bych *et al.* 2008; Sheftel *et al.* 2009) – which was not identified in the original paper. However, it missed one complex I associated gene (*FOXRED1*) which was identified in the original paper, and which has since been confirmed as a complex I assembly factor (Formosa *et al.* 2015). *FOXRED1* did not fulfil the phylogenetic profiling criteria in this analysis, as orthologues were not predicted in some of the designated yeast species known to encode complex I (Figure 3.1). This may be the result of the more specific orthology prediction methodology I used incorrectly removing some true orthologues of *FOXRED1*.



Figure 3.1 (overleaf) Phylogenetic profile of known complex I subunits, associated genes and predicted genes. Each box represents presence or absence of a single gene in a species or taxon. A filled box represents the prediction of an orthologue of the labelled gene in at least one species of the labelled taxon. Numbers in brackets indicate the number of species represented by the label. Species or taxa labelled in red are known to be missing complex I. Gene labels are coloured if they match at least one of the investigated complex I phylogenetic profiles: (i) the COPP1 profile from the MitoCarta paper, (ii) the COPP1B profile, which removes the requirement for a bacterial orthologue, and (iii) the full complex I phylogenetic profile, which requires gene loss from several additional species lacking complex I not studied in the MitoCarta paper. Approximately half of the known complex I associated genes are not predicted by any tested phylogenetic profile. Predicted genes are IMPI genes not currently associated with complex I, which fit at least one of the studied phylogenetic profiles. No predicted genes fit the full complex I phylogenetic profile. (Green = match the COPP1 profile and the full complex I phylogenetic profile; yellow = match the COPP1 profile but not the full complex I phylogenetic profile; blue = match the COPP1B profile and the full complex I phylogenetic profile; orange = match the COPP1B profile but not the full complex I phylogenetic profile; pink = gene identified in the MitoCarta paper, but not this analysis).

I then extended the analysis by removing the criteria for a bacterial orthologue, due to the high proportion of genes with no bacterial orthologues in my dataset (75%) compared to the original paper dataset (25%). The seventeen additional genes that fit these criteria were designated ‘COPP Tier 1B’ (COPP1B) genes and consisted of nine known complex I subunits, four known complex I assembly factors and four other genes (highlighted in blue and orange in *Figure 3.1*). These genes were all identified as fitting the COPP1 profile in the original paper (Pagliarini *et al.* 2008).

The wide range of species included in my orthology dataset provided an opportunity to make a more specific analysis, by including criteria for twenty-four additional species known to be missing complex I which had not been included in the original analysis. These included members of entirely different phyla (the Parabasalia, Chromerida and Dinoflagellata). I refer to this as the ‘full complex I phylogenetic profile’. While these additional criteria did not identify any new genes potentially associated with complex I activity, it did rule out all the genes previously predicted as COPP1 or COPP1B genes which have not already been characterised as complex I associated (genes ruled out by the full complex I phylogenetic profile are highlighted in yellow and orange in *Figure 3.1*). Thus, whilst the expansion of the criteria did not identify any potential new associated genes, it did seemingly correctly rule out

several false predictions, showing the power of the increased scope of the phylogenetic analysis.

One of these specific changes in gene identification compared to the original analysis was the exclusion of *LYRM5* as a gene with a function potentially associated with complex I. This occurred due to better separation of the LYRM family using the reciprocal best hit technique, rather than the simple one-way BLAST. *LYRM5* has recently been identified as a regulatory protein of the electron transfer flavoprotein, rather than a complex I associated protein (Floyd *et al.* 2016), validating this change.

Electron transfer flavoprotein (ETF)

While *LYRM5* no longer fit the pattern for a complex I associated protein, it did fit the observed pattern of loss for the core subunits of the electron transfer flavoprotein (ETF) (Figure 3.2). As I now had the pattern of loss of the ETF across the species in the orthology dataset, I used this to identify additional genes with possible electron transfer flavoprotein associated functions. I combined the pattern of loss across the studied species with criteria specifying identification in at least two free-living cnidarians and two of the remaining platyhelminthes – this identified thirty potentially associated genes. Removing genes with already well-defined functions left six genes (Figure 3.2).

Of these six genes, *C8orf82* is of particular interest. *C8orf82* is one of only five genes with more than one experimental association with *ETF*A in the BioGRID database (Islamaj Doğan *et al.* 2017). (Two of the others are *LYRM5* and *ETFB*, which are characterised parts of the electron transfer flavoprotein or known to affect its function.) Both of these interactions are from high throughput affinity-capture mass spectrometry experiments (Huttlin *et al.* 2015, 2017). *C8orf82* was included in the IMPI positive training set due to strong evidence from an ascorbate peroxidase (APEX) tagging experiment (Hung *et al.* 2014) and contains a domain of unknown function (DUF4504) – the only protein in the human genome to contain this domain. *C8orf82* is well expressed in most human tissues, as are *ETF*A and *ETFB* (Uhlén *et al.* 2015). Therefore, phylogenetic profiling suggests that *C8orf82* would be a candidate for functional studies related to electron transfer flavoprotein, especially given how little else is known about its function.

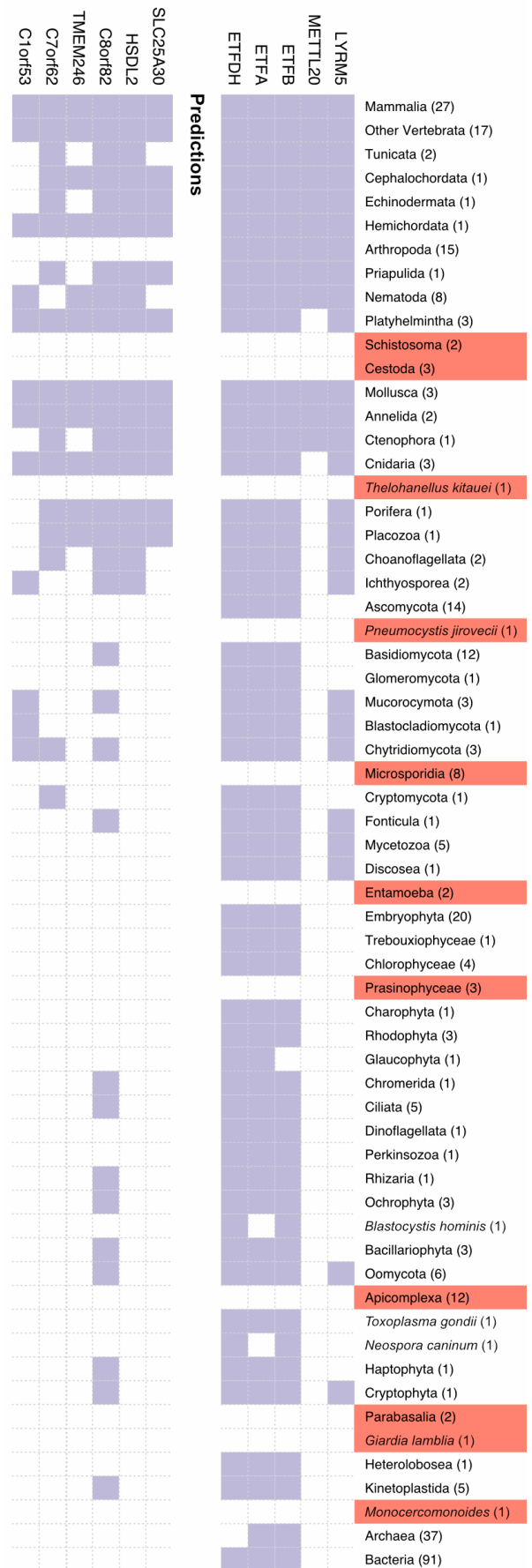


Figure 3.2 Summary phylogenetic profile of genes encoding known electron transfer flavoprotein (ETF) subunits and regulatory proteins, as well as possible associated genes which fit the phylogenetic pattern. Each box represents presence or absence of a single gene in a species or taxon. A filled box represents the prediction of an orthologue of the labelled gene in at least one species of the labelled taxon. Numbers in brackets indicate the number of species represented by the label. Species or taxa labelled in red are missing ETF.

Other complexes and the limitations of phylogenetic profiling

Attempting this phylogenetic profiling analysis helped identify restrictions which limit the utility of this technique – particularly for its application in the understanding of human genes. One of the limitations of phylogenetic profiling in this context is that it is restricted by the identification of species which show variation in the pathway or complex of interest. So, in the case of the complex I analysis, the closest studied species to humans with a loss of complex I were members of the fungi, such as *S. cerevisiae*. Whilst there was variation in the actual identified complex I subunits within the studied holozoan species, there was no holozoan species studied with a complete loss of complex I. This means that the possible function of any genes which originate more recently than the fungi would not be addressed by this methodology. Seeing as nearly 45.5% of the IMPI 2017 genes were only identified in holozoan species (see *Chapter 2*), this is a considerable limitation. Therefore, phylogenetic profiling for the purpose of identifying the potential functions of human genes will potentially be most useful for complexes or pathways with patterns of loss in species which share a more recent common ancestor with humans.

Another limitation is that, for well-conserved pathways or complexes, species (often parasites) which have lost one complex or pathway have lost many others, which makes it much harder to predict the potential function of genes fitting the same phylogenetic profile. A good example of this is seen in the phylogenetic profiles of the respiratory complexes II-IV and ATP synthase (*Figures 3.3-3.6*). Species that have lost one of these complexes have nearly always lost all of them. Not only that, but these species are all known to contain degenerate forms of mitochondria (mitosomes or hydrogenosomes), so also do not encode many genes necessary for other parts of mitochondrial function or metabolism, with only between 4.7% and 13.2% of IMPI 2017 genes identified in these species. This makes it difficult to predict the potential function of the genes which have been lost in this species, as they could function in any number of pathways or complexes. Complexes or pathways considered for phylogenetic profiling should therefore preferably show good variation of presence and absence across a variety of species, and not just those which have highly divergent or minimal genomes.

Despite the limitations of the technique, this analysis confirms that phylogenetic profiling is a potentially useful tool in predicting the function of human genes. The limitations of the technique should be considered before use, to identify good potential targets for study.

Parasitic species within the Holozoa, such as the myxozoans and some nematodes and platyhelminthes, may be good species to consider for phylogenetic profiling. They share a more recent common ancestor with humans (in comparison to *Saccharomyces cerevisiae*) and their genomes contain up to 48% of IMPI 2017 genes, increasing the potential number of genes which can be analysed.

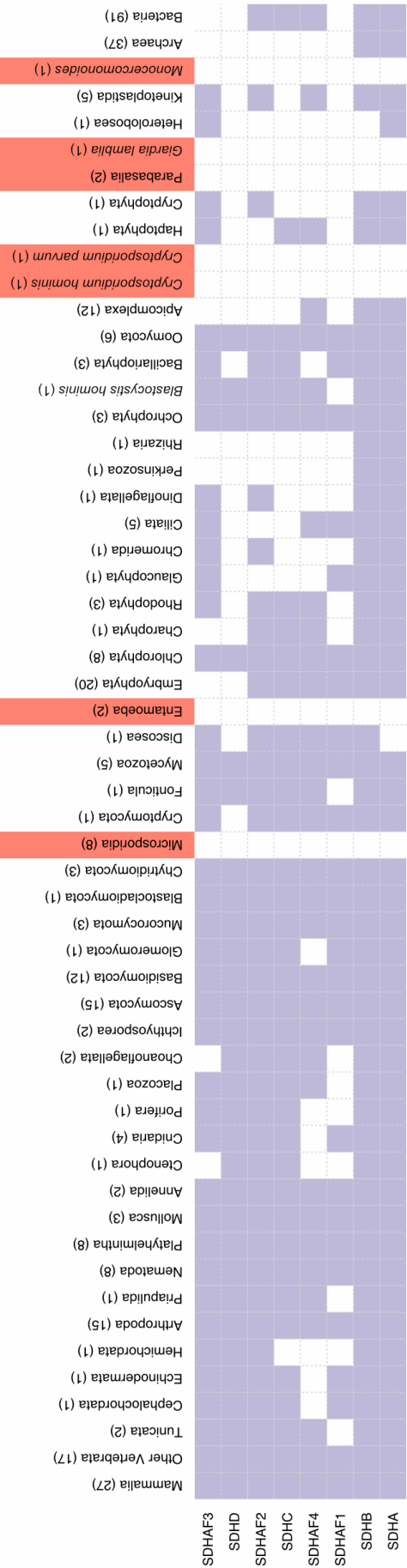
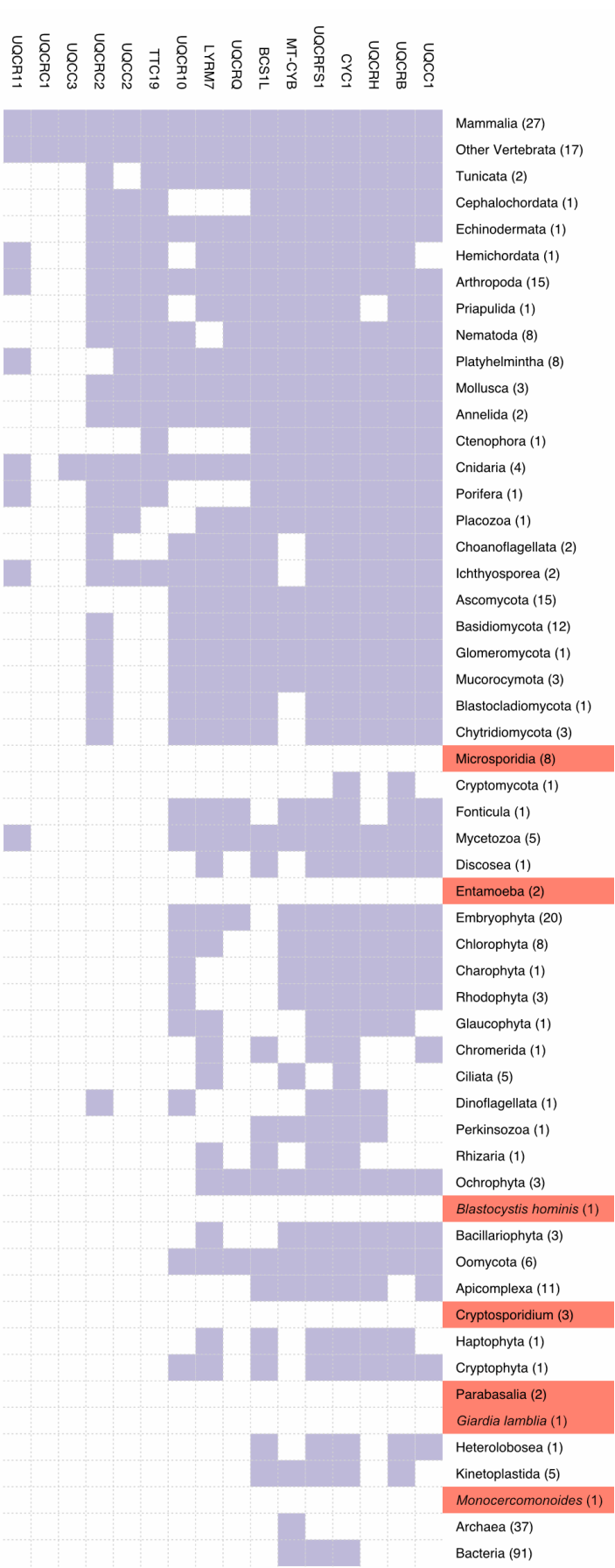


Figure 3.3 Summary phylogenetic profile of known complex II subunits and assembly factors. Each box represents presence or absence of a single gene in a species or taxon. A filled box represents the prediction of an orthologue of the labelled gene in at least one species of the labelled taxon. Numbers in brackets indicate the number of species represented by the label. Species or taxa labelled in red are missing complex II.

Figure 3.4 Summary phylogenetic profile of known complex III subunits and assembly factors. Each box represents presence or absence of a single gene in a species or taxon. A filled box represents the prediction of an orthologue of the labelled gene in at least one species of the labelled taxon. Numbers in brackets indicate the number of species represented by the label. Species or taxa labelled in red are missing complex III.



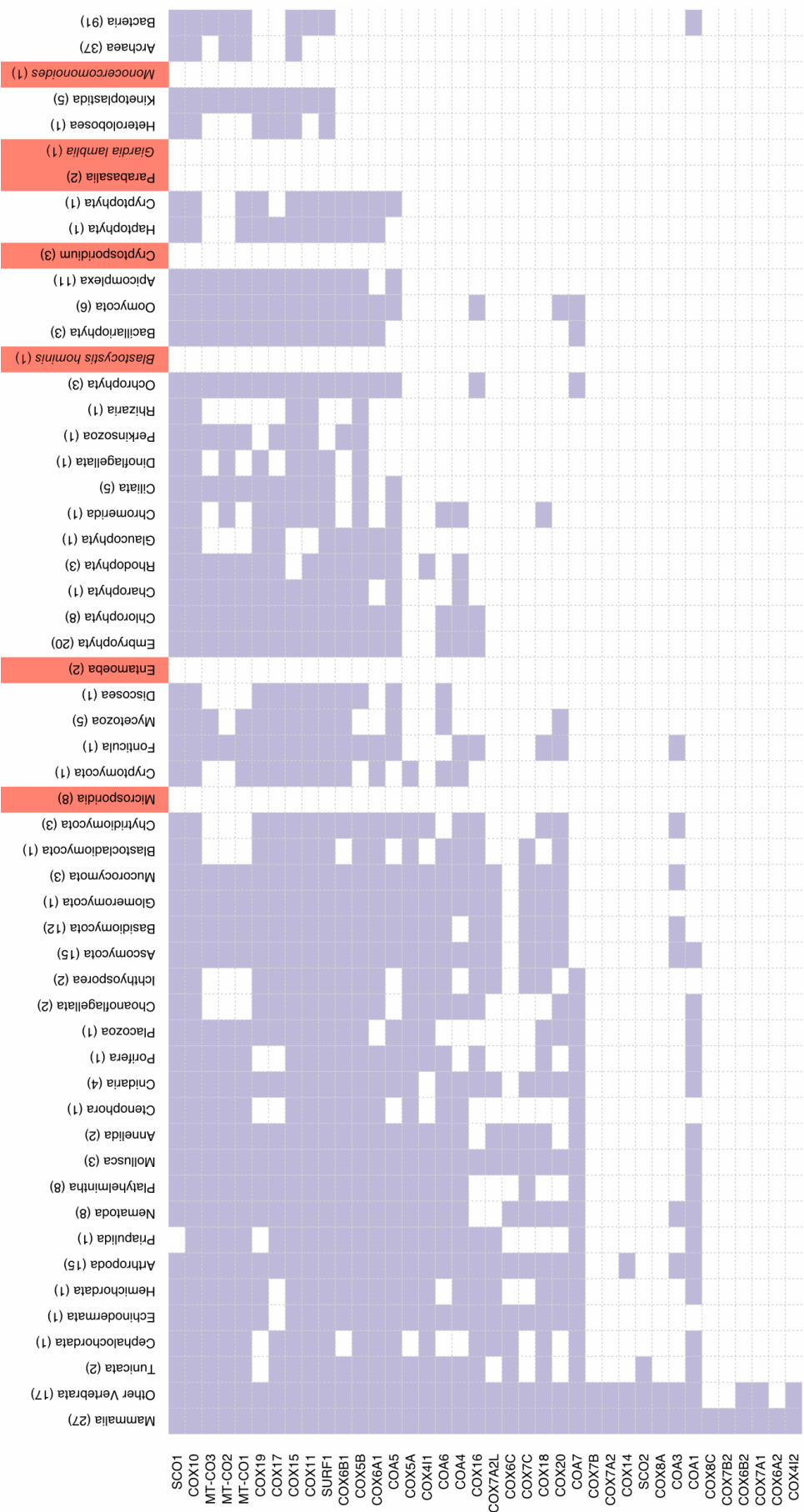
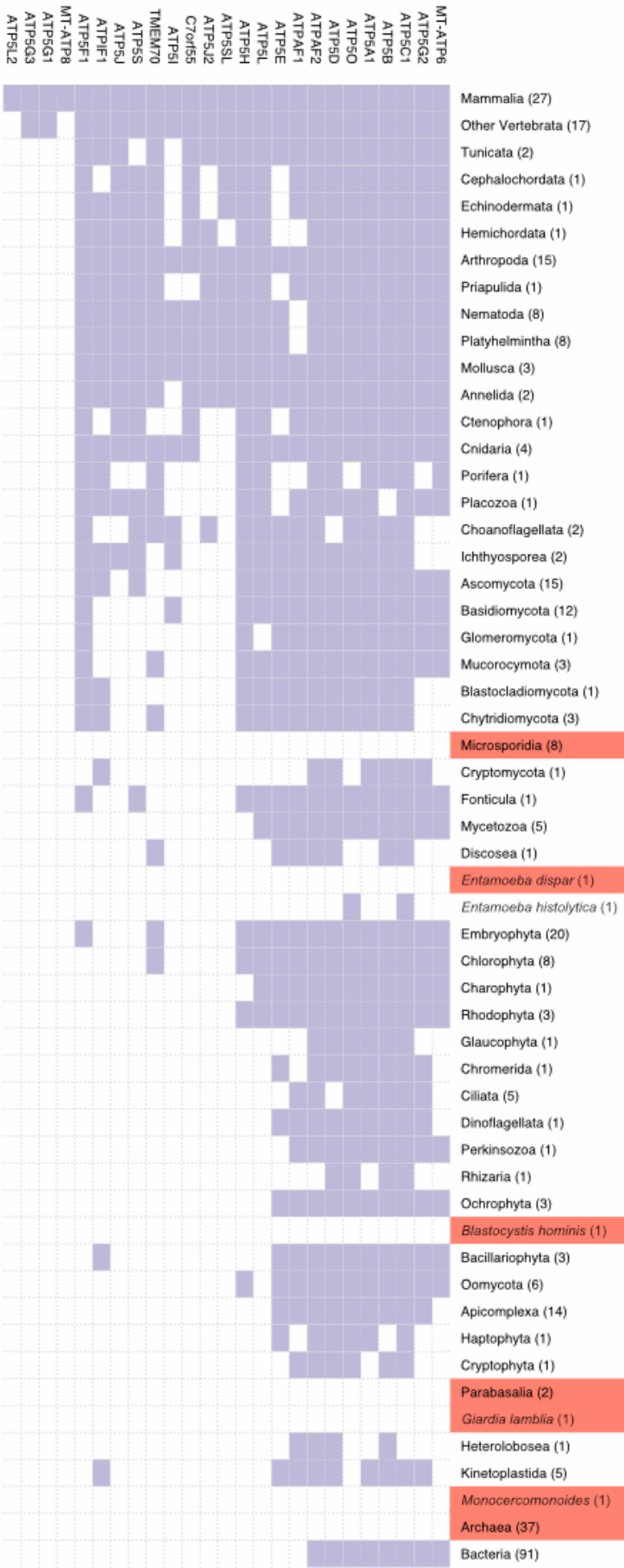


Figure 3.5 Summary phylogenetic profile of known complex IV subunits and assembly factors. Each box represents presence or absence of a single gene in a species or taxon. A filled box represents the prediction of an orthologue of the labelled gene in at least one species of the labelled taxon. Numbers in brackets indicate the number of species represented by the label. Species or taxa labelled in red are missing complex IV.

Figure 3.6 Phylogenetic profile of known ATP synthase subunits and assembly factors. Each box represents presence or absence of a single gene in a species or taxon. A filled box represents the prediction of an orthologue of the labelled gene in at least one species of the labelled taxon. Numbers in brackets indicate the number of species represented by the label. Species or taxa labelled in red are missing ATP synthase.



Conclusions

In this chapter, I have described a phylogenetic profiling analysis based around the orthology dataset created in *Chapter 2*. This analysis attempts to predict potential assembly factors of the mitochondrial respiratory complexes, particularly complex I and electron transfer flavoprotein. The analysis did not identify any new potential complex I assembly factors, though the increased phylogenetic spread and different orthology detection method improved the accuracy of the identification of the known complex I subunits and assembly factors, by removing several false positives. Complex I assembly factors discovered in the future are more likely to be novel to the Holozoa. The analysis predicted the association of the uncharacterised gene *C8orf82* with the electron transport flavoprotein, which is supported by some literature evidence. Further work is necessary to confirm this prediction. More generally, I have derived guidelines for features of species, complexes and pathways which will be most amenable to future phylogenetic profiling attempts.

Chapter 4

Function of the mitochondrial carriers

Introduction

Lessons from Chapter 3

Whilst phylogenetic profiling is a powerful technique for predicting gene function, the analyses I carried out in *Chapter 3* using this technique to explore the mitochondrial respiratory complexes gave indications of some limitations of the methodology. The success of this technique is dependent on finding genes with strong phylogenetic profiles, preferably in species which share a relatively recent common ancestor with the original species of interest (in this case humans). In general, parasitic species are a good source of variation in phylogenetic profiles, as their dependence on hosts for a variety of metabolic processes often leads to comparative gene loss (Keeling & Slamovits 2005). However, parasites which have lost too much metabolism become less useful.

The orthology dataset from *Chapter 2* was useful as a source of information on which families of genes may be tractable to phylogenetic profiling, within certain taxa. These could then be taken forward for a further targeted analysis. It was noticeable that there was variation in the presence and absence of different members of the mitochondrial carrier gene family across eukaryotic species, with the differences present even within the Metazoa.

Mitochondrial carriers

Mitochondria form distinct compartments within the cell, bounded by a double membrane. The outer mitochondrial membrane is relatively porous, due to the presence of porin structures (*VDAC1*, *VDAC2* and *VDAC3* in humans). These β -barrel proteins form channels in the outer mitochondrial membrane, allowing movement of molecules less than approximately 5,000 Da between the cytoplasm and the intermembrane space (Sorgato *et al.* 1993). Passage of molecules across the inner mitochondrial membrane is more tightly controlled. This means the cell can control the localisation and concentrations of a variety of molecules and metabolites, allowing mitochondrial compartmentalisation of certain processes and the formation of concentration gradients within the cell (e.g. the H^+ gradient across the inner mitochondrial membrane).

The mitochondrial carrier family of proteins form a large group of transporters, mostly localised to the inner mitochondrial membrane. There are fifty-three members of this family in humans (Palmieri 2013). The basic structure of proteins within the carrier family is three homologous repeats of approximately 100 aa each (Walker & Runswick 1993). Each repeat consists of two transmembrane helices (for a total of six in the protein) (*Figure 4.1*), and they are linked by shorter α -helices on the matrix side (Pebay-Peyroula *et al.* 2003).

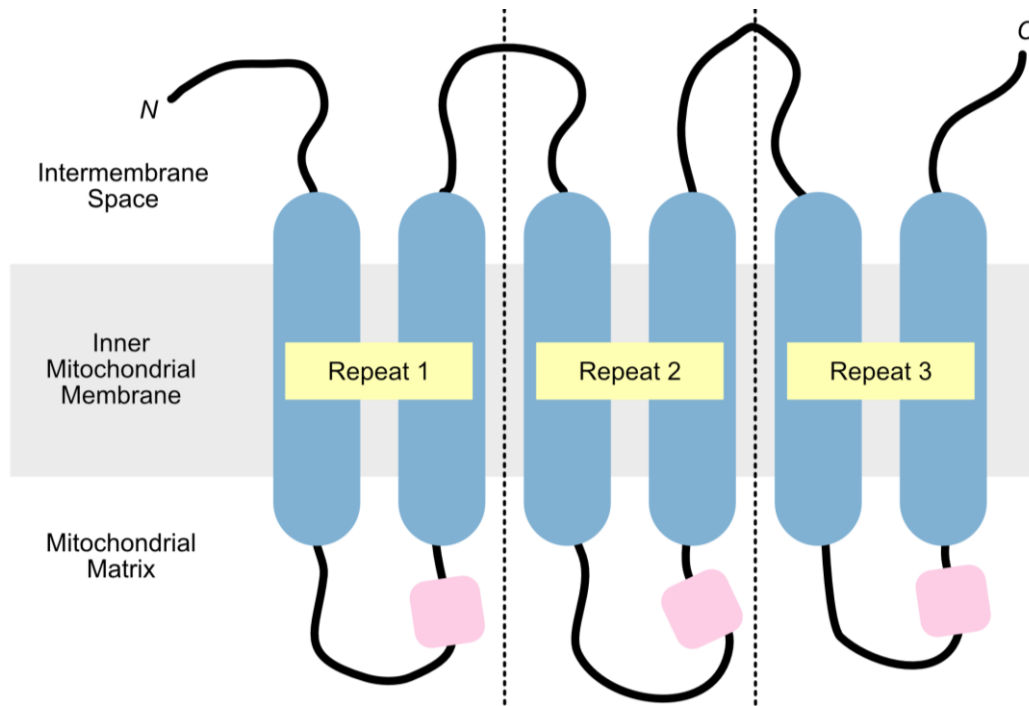


Figure 4.1 General structure of proteins in the mitochondrial carrier family. Six transmembrane helices (blue) are ordered in a three-fold homologous structure. Short alpha helices (pink) connect the helices on the matrix side.

Evidence suggests that transport of substrates by members of this family occurs through movement of the protein between two conformational states – one which is open to the intermembrane space, and one which is open to the matrix (Kunji *et al.* 2016). Each of the three repeats in the carrier contains two signature motifs: PX[DE]XX[RK] on the matrix side (Nelson *et al.* 1998) and [FY][DE]XX[RK] on the intermembrane space side (Robinson *et al.* 2008). The charged residues can form salt bridges, which allows the transporter to be held open to either the matrix side or the intermembrane space side, whilst the opposite side is closed off, facilitating the transport process (King *et al.* 2016).

Transport function for several mitochondrial carriers has been characterised. Members of this family transport a wide range of substrates, including nucleotides, amino/keto acids and

inorganic ions, such as iron and phosphate (Gutiérrez-Aguilar & Baines 2013). Other members function in processes beyond transport activity. For example, *UCP1* is linked to the production of heat in brown adipose tissue, which is important in keeping new-born babies warm, through the uncoupling of proton movement across the inner mitochondrial membrane from ATP synthesis (Nedergaard *et al.* 2001). *MTCH1* and *MTCH2* are localised to the outer mitochondrial membrane and are associated with apoptosis (Mao *et al.* 2008; Robinson *et al.* 2012). However, many members of the mitochondrial carrier family await functional characterisation, providing an opportunity for bioinformatic input.

Chapter summary

In this chapter, I identify phyla with variation in mitochondrial carrier presence and absence which support further study through phylogenetic profiling. I build a dataset of orthologues of genes encoding the mitochondrial proteome within two phyla – the Nematoda and the Platyhelmintha. I use this dataset to attempt to match phylogenetic profiles of characterised mitochondrial carriers to missing metabolic genes, to assess the likelihood of success of this methodology and identify features which may be indicative of a link to function. I then use what I have learnt to explore the possible transport activities of uncharacterised members of the mitochondrial carrier family in humans.

Methods

Human carrier phylogenetic tree

To create the tree of human mitochondrial carrier family proteins, I retrieved the canonical human protein sequences for each carrier from the NCBI (National Center for Biotechnology Information) protein database (<https://www.ncbi.nlm.nih.gov/protein>). These sequences were aligned using MUSCLE with default settings (Edgar 2004a, 2004b) and alignments were manually assessed in Jalview (Waterhouse *et al.* 2009). A phylogenetic tree was calculated using PhyML 3.0 using the best method (highest maximum likelihood) from the two topology search methods (SPR – Subtree Pruning and Regrafting; and NNI – Nearest Neighbour Interchange) and 100 bootstrap replicates. The tree was visualised in Interactive Tree Of Life (Letunic & Bork 2016).

Identifying carriers by using sequence clustering and hidden Markov models

Identification and classification of orthologues of the mitochondrial carriers in platyhelminthes and nematodes was carried out by Dr Alan Robinson of the Bioinformatics group, Medical Research Council Mitochondrial Biology Unit, University of Cambridge, CB2 0XY, UK.

Proteomes of representative multicellular metazoan species having a fully sequenced and annotated genome were downloaded, including chordates (*Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Mus musculus*, *Rattus norvegicus*, *Canis familiaris*, *Monodelphis domestica*, *Gallus gallus*, *Taeniopygia guttata*, *Xenopus tropicalis*, *Danio rerio*, *Fugu rubripes*, *Tetraodon nigroviridis*, *Branchiostoma floridae*, *Ciona intestinalis*, *Oikopleura dioica*, *Halocynthia roretzi*), arthropods (*Drosophila melanogaster*, *Drosophila pseudoobscura*, *Pediculus humanus corporis*, *Ixodes scapularis*), nematodes (*Loa loa*, *Caenorhabditis briggsae*, *Caenorhabditis elegans*), platyhelminthes (*Clonorchis sinensis*, *Echinococcus multilocularis*, *Echinococcus granulosus*, *Hymenolepis microstoma*, *Opisthorchis viverrini*, *Schistocephalus solidus*, *Schistosoma japonicum*, *Schistosoma haematobium*, *Schistosoma mansoni*), cnidaria (*Hydra magnipapillata*, *Nematostella vectensis*), and porifera (*Amphimedon queenslandica*).

Sequences of mitochondrial carriers in each species' proteome were identified by using the Pfam model for the mitochondrial carrier family (Pfam: MitoCarr) and by BLASTp searches (Altschul *et al.* 1990, 1997) with human mitochondrial carriers. First, the chordate carriers were grouped and aligned into subfamilies using clustering with ClustalW (Thompson *et al.* 2002). Then, sequences from phyla progressively more distant from chordates were assigned to subfamilies by using BLASTp searches together with ClustalW alignments and clustering. A sequence was assigned to a new subfamily if it was not a reciprocal best hit of an existing subfamily. Finally, sequences were aligned by using MUSCLE (Edgar 2004a, 2004b) and Jalview (Waterhouse *et al.* 2009), and inspected manually. Each subfamily contained one sequence per species, except in the case of indistinguishable paralogs (e.g. *UCP1*, *UCP2* and *UCP3*).

The multiple sequence alignment of each subfamily was used as a seed alignment to calculate a hidden Markov model (HMM) by using HMMER (Eddy 1998).

Protein sequences from nematodes and platyhelminthes having more than 1,000 protein sequences in the NCBI 'nr' database were downloaded, as summarised in *Table 4.1*. Sequences for the free-living platyhelminth *Schmidtea mediterranea* were also retrieved from WormBase ParaSite (Howe *et al.* 2017). These protein sequences were searched for mitochondrial carriers using HMMER (Eddy 1998) and the HMM of the mitochondrial carrier sub-families generated previously (E-value threshold of 1.0). Each protein sequence was initially assigned to the subfamily with which it had the most significant HMM match. Sequences were aligned to the original seed subfamily by using MUSCLE (Edgar 2004a, 2004b) and Jalview (Waterhouse *et al.* 2009). Each subfamily's alignment was inspected manually to check for contaminants and for sequences of new subfamilies.

Building nematode and platyhelminth orthologue datasets

Reciprocal best hit was used to identify orthologues of proteins encoded by the human IMPI positive training set genes in ten platyhelminthes and twenty-five nematodes (Table 4.1).

Table 4.1 Species of platyhelminthes and nematodes included in the phylogenetic profiling analysis.

Phyla		Species
Platyhelmintha	Tricladida	<i>Schmidtea mediterranea</i>
	Plagiorchiida	<i>Clonorchis sinensis</i>
	Plagiorchiida	<i>Opisthorchis viverrini</i>
	Schistosoma	<i>Schistosoma mansoni</i>
	Schistosoma	<i>Schistosoma japonicum</i>
	Schistosoma	<i>Schistosoma haematobium</i>
	Cestoda	<i>Echinococcus granulosus</i>
	Cestoda	<i>Echinococcus multilocularis</i>
	Cestoda	<i>Hymenolepis microstoma</i>
	Cestoda	<i>Taenia solium</i>
Nematoda	Diplogasterida	<i>Pristionchus pacificus</i>
	Rhabditida	<i>Caenorhabditis briggsae</i>
	Rhabditida	<i>Caenorhabditis elegans</i>
	Rhabditida	<i>Caenorhabditis remanei</i>
	Rhabditida	<i>Caenorhabditis brenneri</i>
	Rhabditida	<i>Haemonchus contortus</i>
	Rhabditida	<i>Necator americanus</i>
	Rhabditida	<i>Ancylostoma ceylanicum</i>
	Rhabditida	<i>Ancylostoma duodenale</i>
	Rhabditida	<i>Oesophagostomum dentatum</i>
	Ascaridida	<i>Ascaris suum</i>
	Ascaridida	<i>Toxocara canis</i>
	Filarioidea	<i>Brugia malayi</i>
	Filarioidea	<i>Wuchereria bancrofti</i>
	Filarioidea	<i>Loa loa</i>
	Dorylaimia	<i>Trichinella britovi</i>
	Dorylaimia	<i>Trichinella murrelli</i>
	Dorylaimia	<i>Trichinella native</i>
	Dorylaimia	<i>Trichinella nelson</i>
	Dorylaimia	<i>Trichinella papuae</i>
	Dorylaimia	<i>Trichinella patagonensis</i>
	Dorylaimia	<i>Trichinella pseudospiralis</i>
	Dorylaimia	<i>Trichinella spiralis</i>
	Dorylaimia	<i>Trichinella zimbabwensis</i>
	Dorylaimia	<i>Trichuris suis</i>

Two BLASTp protein sequence databases were built – one containing sequences from the platyhelminth species and one containing sequences from the nematode species. Any identical protein sequences within each species were removed. For each IMPI positive training set gene, a BLASTp search was performed against each database, using the canonical human protein sequence as a bait sequence (E-value cut-off 1×10^{-5}). For each species, the top hit(s) for each gene was chosen for a reciprocal BLASTp search against a dataset of human proteins (E-value cut-off 1×10^{-5}). Each sequence which returned a human protein equivalent to the original human gene as the top hit in the reciprocal BLASTp search was called an orthologue (summarised in *Figure 4.2*).

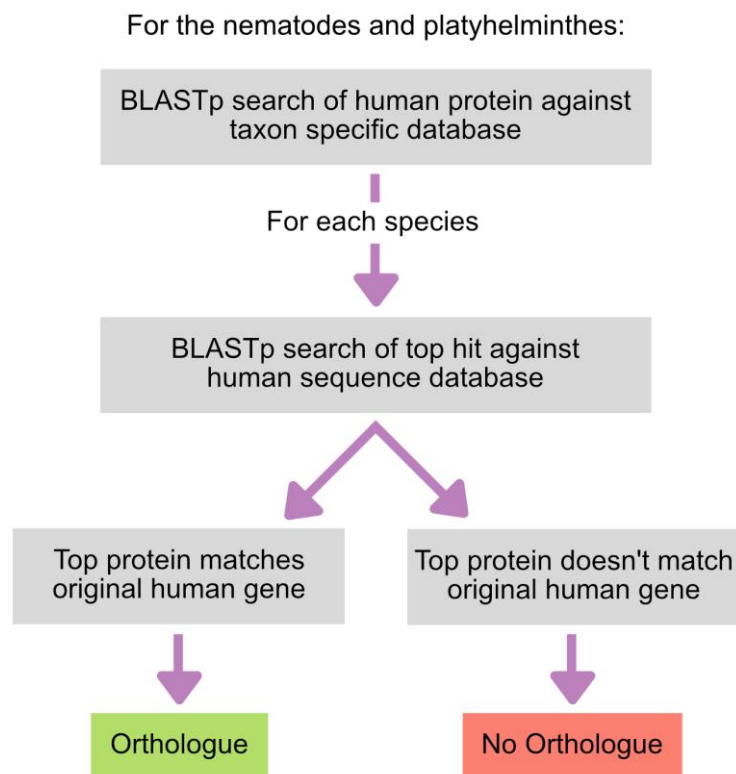


Figure 4.2 Summary of the reciprocal best hit method for identifying orthologues of IMPI 2017 positive training set proteins in the targeted nematodes and platyhelminthes.

Manual analysis was then used to attempt to fill any gaps in orthologue identification and to clarify assignments of orthologues to different members of paralogous groups. Additional BLASTp searches were used to identify fragments of proteins missed by the original searches, using the WormBase ParaSite database (Howe *et al.* 2017).

Where there was confusion between the assignment of orthologues between human genes with closely related sequences, phylogenetic trees were used to provide clarification where

possible. Sequences from a variety of model organisms and predicted orthologous sequences from platyhelminthes or nematodes were aligned using MUSCLE with default settings (Edgar 2004a, 2004b). PhyML 3.0 was used to calculate a predicted phylogenetic tree for these sequences using the best method (highest maximum likelihood) from the two available topology search methods and 100 bootstrap replicates (Guindon *et al.* 2010). Manual assessment of the trees was used to clarify assignments of nematode and platyhelminth sequences to an appropriate human gene. Efforts were made to maintain consistency for assignment decisions made within the studied phyla and between the platyhelminth and nematode analyses.

Clustering

R was used to cluster genes by their phylogenetic patterns. Distance matrices were calculated from the phylogenetic pattern matrix using the function *dist()*, with the method ‘*manhattan*’. Clustering was achieved using the function *hclust()*, with the method ‘*ward.D2*’.

Pathway analysis

Pathway enrichment analysis was carried out in Reactome (Fabregat *et al.* 2018), using the ‘Analyze Pathway’ tool. This tool looks for Reactome pathways which are significantly overrepresented in the given list of genes and is based on known human genes and pathways.

Mitochondrial targeting sequence prediction

Four programmes were used to produce mitochondrial targeting sequence predictions:

- iPSORT (Bannai *et al.* 2002) – which returns a binary prediction (0 or 1),
- MitoProt (Claros & Vincens 1996)– which returns a score between 0 and 1,
- TargetP (Emanuelsson *et al.* 2000) – which returns a score between 0 and 1,
- MitoFates (Fukasawa *et al.* 2015) – which returns a score between 0 and 1.

For the continuous prediction scores, predictions were considered of interest if the score was above 0.75.

Human mitochondrial carrier family

80

This tree shows a distinct pattern of some highly supported groups of genes (marked in grey), joined at a poorly supportive centre. All characterised genes group together with others of the same function. For example, the four adenine nucleotide transporters (*SLC25A4/5/6/31*) group together with a completely supportive bootstrap value (100/100). I therefore used this tree to identify three groups of the genes without characterised functions, whose patterns of presence and absence should be combined when looking at phylogenetic profiling: *SLC25A45/47/48*, *SLC25A14/30*, and *SLC25A51/52*. These groups were taken forward to the next stage of the analysis.

Identifying useful species for phylogenetic profiling

One potential way to characterise the function of the carrier proteins is to match patterns of missing metabolism and missing carriers across different species. Previous work looking at the respiratory complexes (*Chapter 2*) indicated that the best groups of species to study should fulfil several criteria. First, the date of the most recent common ancestor of the investigated taxon and the original species of interest (human) should be as recent as possible, as the more recently the two taxa share a common ancestor, the more likely the metabolic pathways are similar. Additionally, using metazoan species means they are more likely to contain pathways which are unique to the Metazoa, which is important as a large proportion of human mitochondrial proteins only had predicted orthologues within the Metazoa (*Chapter 2*). Secondly, species within the taxon should have varied patterns of gene loss, which may be exemplified by species with different types of parasitic lifestyles. Thirdly, any taxon with studied parasitic species should contain at least one less-degenerate (e.g. free-living) species to compare to.

Two different phyla fit these criteria: the Nematoda (nematodes or roundworms) and the Platyhelmintha (platyhelminthes or flatworms). Both phyla contain multicellular animals, some of which are parasitic with varying types of parasitic lifestyles and some of which are free-living. Reasonable numbers of each phyla have had their genome sequenced to allow comparative analysis.

Hidden Markov models (HMMs) were used to identify the presence of mitochondrial carriers in ten species of platyhelminth and twenty-five species of nematode worm with fully sequenced genomes. This included multiple species with similar characteristics where

possible, to reduce the mistakes introduced by less complete sequencing/annotation within individual species. Carriers branching closely together in the tree of human mitochondrial carriers (Figure 4.3) were grouped together.

The platyhelminthes included in this analysis have four different lifestyles: free-living, liver flukes, blood flukes and tapeworms. Nineteen of the mitochondrial carrier groupings were identified in all or nearly all the studied platyhelminthes with four groups identified in none (Figure 4.4).

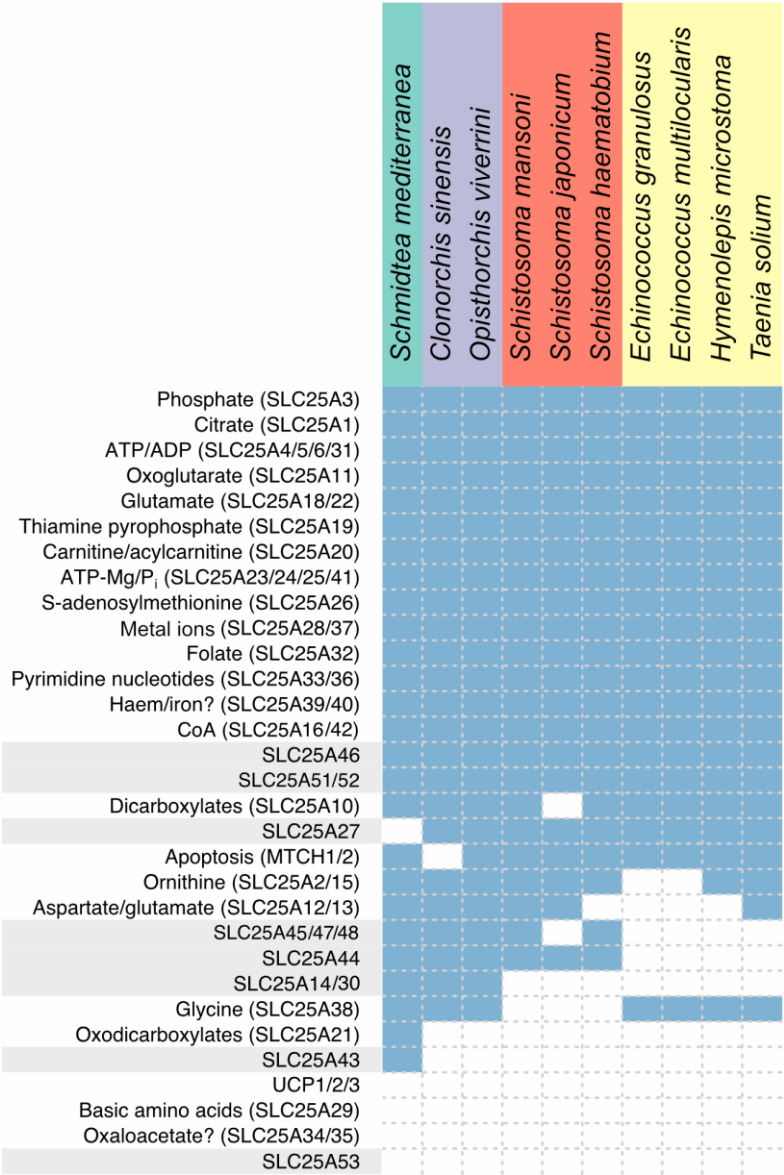


Figure 4.4 Phylogenetic profile of human mitochondrial carrier family proteins in ten platyhelminthes. Boxes filled in blue represent a carrier identified in labelled. Species are colour-coded by lifestyle: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Carrier groups which are not currently characterised are highlighted in grey.

Eight mitochondrial carrier groups show consistent differences in the phylogenetic pattern within the studied platyhelminthes, including four carrier groups with a full or partially characterised function and four with no characterised function. The characterised carriers with differences in phylogenetic patterns provide a good basis for assessing the usefulness of phylogenetic profiling within the platyhelminthes to identify carrier function.

The nematodes included in the analysis are sampled from a wider range of taxonomic groups but, importantly, some are parasitic and some are free-living. Identification of the carriers within the nematodes is patchier than in the platyhelminthes (*Figure 4.5*). Nineteen groups of mitochondrial carriers are identified relatively consistently across the species, with four groups not identified in any studied nematode species. Eight mitochondrial carrier groups show consistent differences between different groups of nematodes, including six with full or partially characterised functions (*Figure 4.5*). The patterns of loss or non-identification of carriers differs between the nematodes and the platyhelminthes, providing two separate data sources to investigate with phylogenetic profiling.

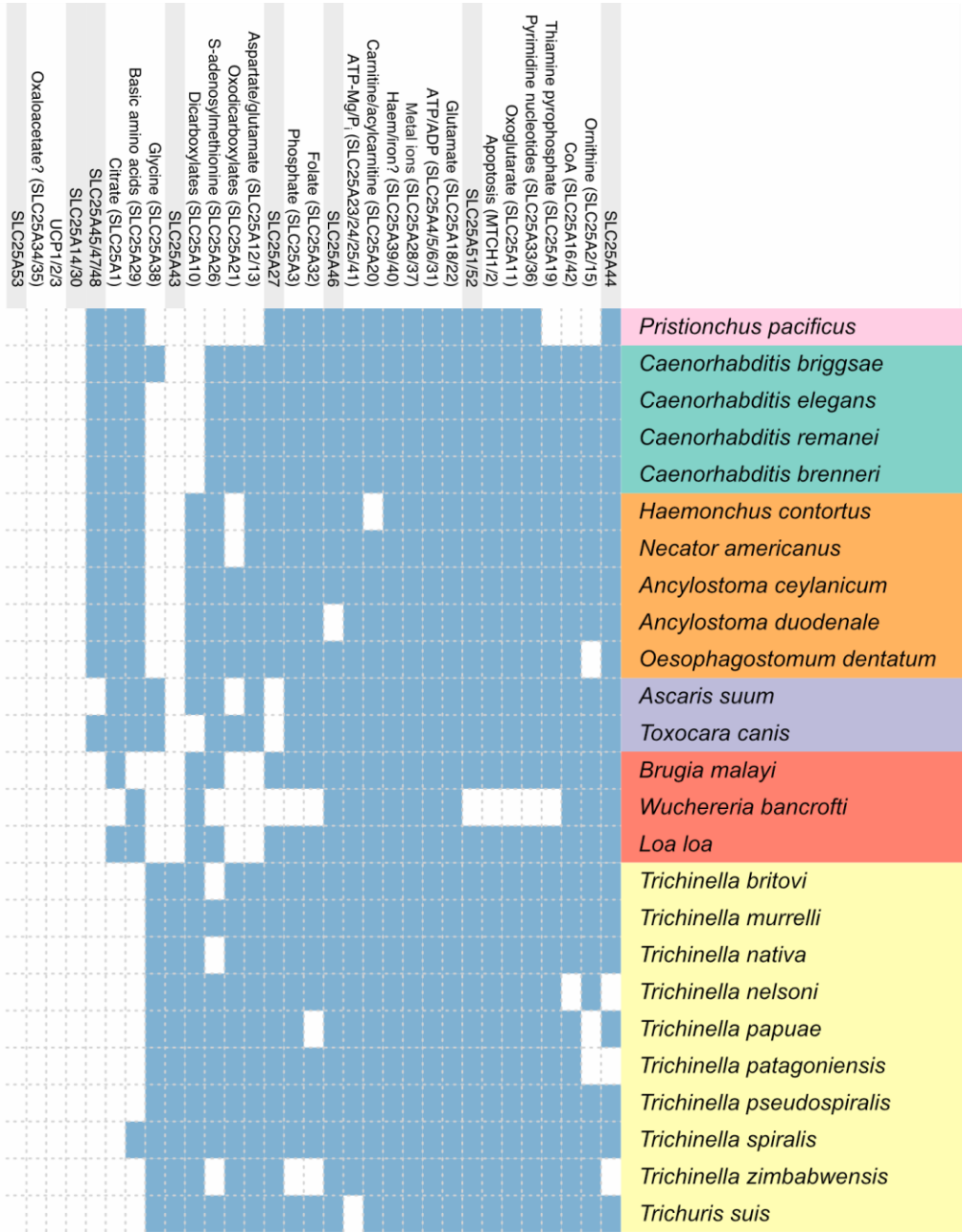


Figure 4.5 Phylogenetic profile of human mitochondrial carrier family proteins in twenty-five nematodes. Boxes filled in blue represent a carrier identified in labelled. Species are colour-coded by lifestyle: pink = free-living *Diplogasterida*; green = free-living *Rhabditida*; orange = parasitic *Rhabditida*; purple = *Ascaridida*; red = *Filarioidea*; yellow = *Dorylaimia*. Carrier groups which are not at all characterised are highlighted in grey.

Building an orthologue dataset

To investigate phylogenetic patterns, I needed a dataset of orthologues of genes within the studied platyhelminthes and nematodes. As the carriers are mainly mitochondrial, I looked only at genes encoding proteins with a mitochondrial function. I decided to use the IMPI positive training set genes for this analysis, which consists of a curated set of 1,130 genes with strong evidence for mitochondrial localisation. The positive training set genes are more certain to have a characterised function within the mitochondria than the predicted genes from the IMPI algorithm. Genes with unknown functions within the mitochondria would not be useful in this analysis, as there would be no metabolic functions to link to transport activity.

Reciprocal best hit analysis identified 4,924 predicted orthologues of 752 IMPI genes in the ten studied platyhelminthes and 10,624 predicted orthologues of 777 IMPI genes in the twenty-five studied nematodes. The more accurate the phylogenetic pattern, the more likely that phylogenetic profiling may be useful in identifying function. I, therefore, carried out two manual steps to improve the completeness and accuracy of orthologue identification. To improve the completeness of the dataset, I carried out manual BLASTp searches using the ParaSite WormBase database (Howe *et al.* 2017) to look for any fragments of orthologous proteins that were potentially missed, for all studied genes with at least one identified orthologue.

To improve the consistency of the orthologue assignment, I looked for groups of human genes with similar sequences and checked for consistent assignment of platyhelminth and nematode orthologues within these groups, using phylogenetic trees where necessary to attempt to identify the closest human orthologue. For example, in the nematode species studied, proteins with FUN14 domains were assigned as orthologues to both *FUNDC1* and *FUNDC2*.

Predicted nematode orthologous sequences cluster together in a phylogenetic tree (*Figure 4.6*), and *FUNDC2* orthologues seem to be slightly less distant from the common ancestral sequence than *FUNDC1* orthologues, so all nematode proteins were assigned as *FUNDC2* orthologues.

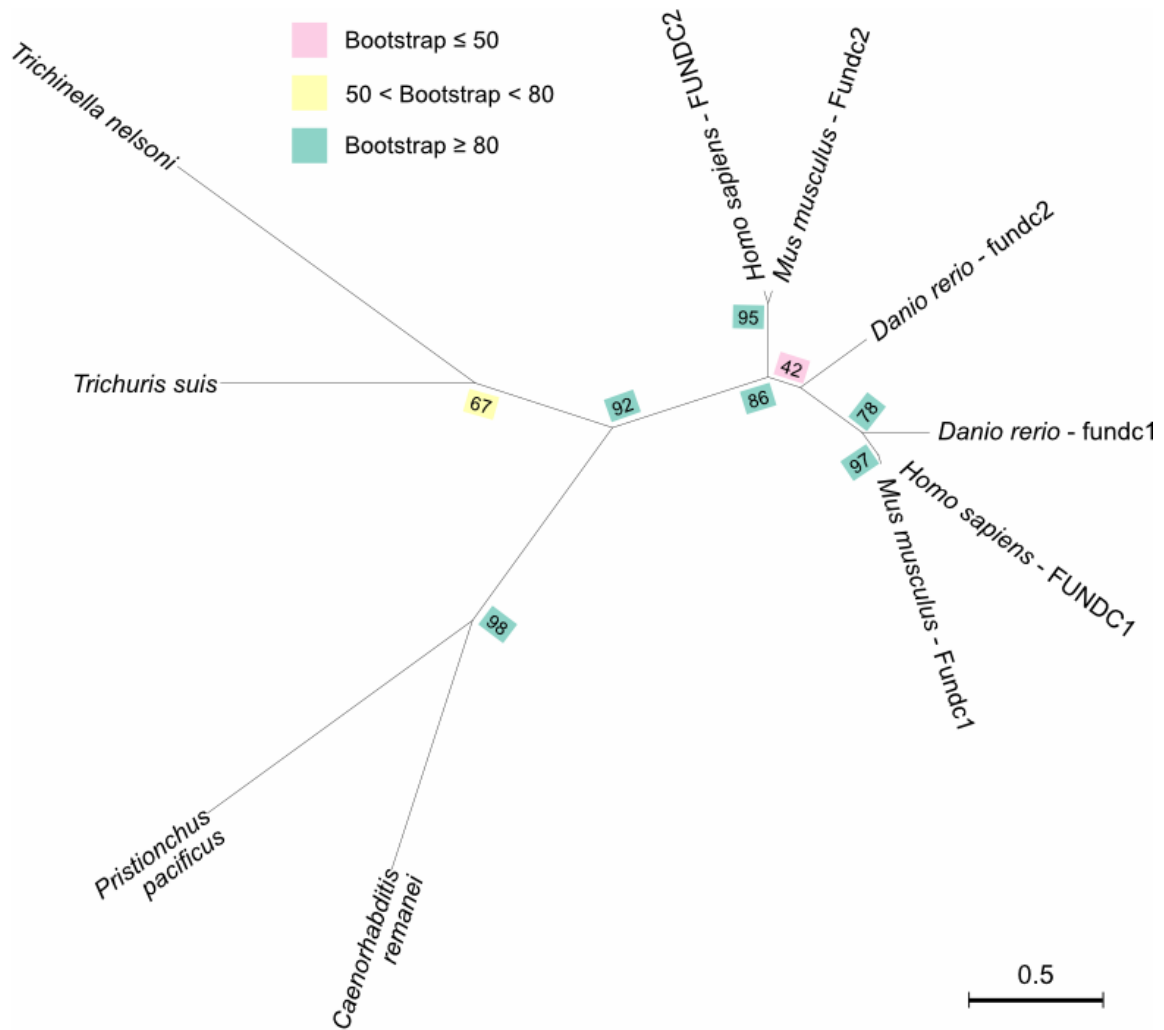


Figure 4.6 Phylogenetic tree of known FUNDC1 and FUNDC2 orthologues from humans and the model organisms *Mus musculus* (mice) and *Danio rerio* (zebrafish), and predicted orthologues from nematodes with bootstrap values (out of 100). Branch length scale indicates number of substitutions per site.

After completion of these steps, the orthologue dataset contains 5,584 predicted orthologues of IMPI positive training genes in the ten platyhelminthes studied; and 10,768 orthologues in the twenty-five nematodes studied. 60.8% (688/1130) of IMPI positive training genes have at least one predicted orthologue in the studied platyhelminthes and 61.5% (696/1130) have at least one predicted orthologue in the studied nematodes. 617 genes have predicted orthologues in both phyla, with 71 genes with orthologues only in the platyhelminthes and 79 genes with orthologues only in the nematodes. Despite both phyla including multicellular members of the Metazoa (animals), 363 of the tested genes (32.1%) do not have any predicted orthologues in either the platyhelminthes or the nematodes. Of the 459 IMPI positive training genes assigned to the Reactome pathway ‘metabolism’, 306 (66.7%) have at least one

identified orthologue in the studied platyhelminthes and 307 (66.9%) have at least one identified orthologue in the studied nematodes, with 256 (55.8%) identified in both phyla.

Whilst most genes have either no predicted orthologues or predicted orthologues for all or nearly all the studied species of each phylum, there are distinct phylogenetic patterns of predicted orthologues in some genes within the platyhelminthes (*Figure 4.7*) and the nematodes (*Figure 4.8*). The different phylogenetic patterns for some genes within the two studied phyla provide a good source of data for attempting phylogenetic profiling to predict the function of the mitochondrial carriers.

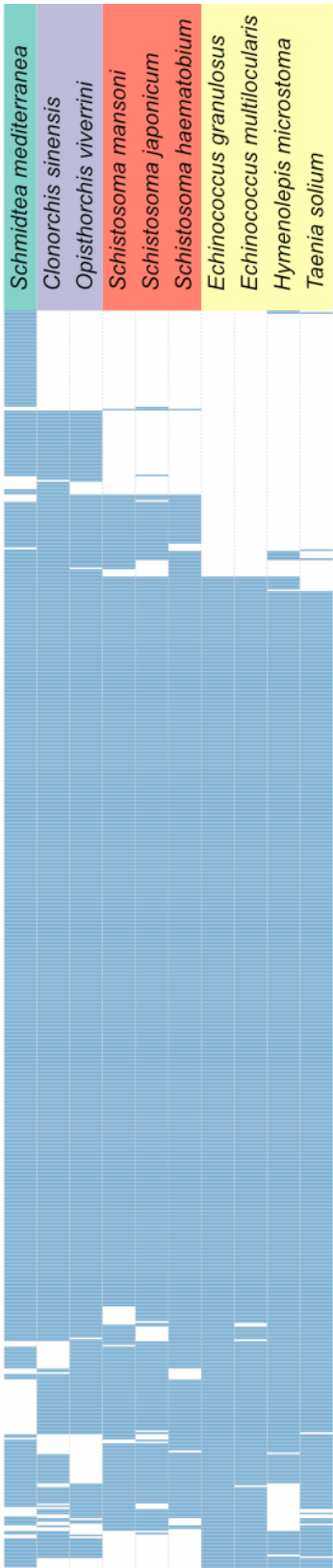


Figure 4.7 Clustered phylogenetic profile of IMPI 2017 positive training genes in ten platyhelminth species. Includes 658 genes with at least one predicted orthologue. Boxes filled in blue represent a carrier identified in labelled. Species are colour-coded by lifestyle: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms.

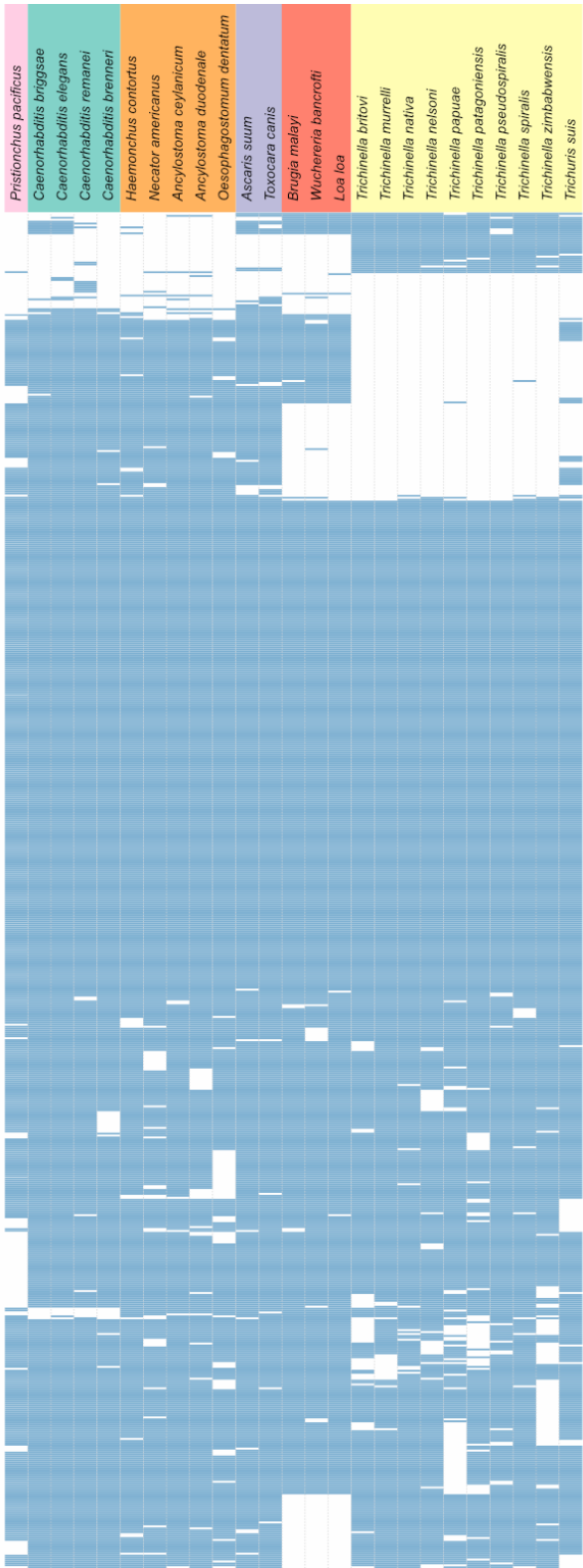


Figure 4.8 Clustered phylogenetic profile of IMPI 2017 positive training genes in twenty-five nematode species. Includes 696 genes with at least one predicted orthologue. Boxes filled in blue represent a carrier identified in labelled. Species are colour-coded by lifestyle: pink = free-living Diplogasterida; green = free-living Rhabditida; orange = parasitic Rhabditida; purple = Ascaridida; red = Filarioidea; yellow = Dorylaimia.

Investigating the characterised transporters

The characterised transporters with distinct phylogenetic patterns provide a source to test whether phylogenetic profiling is feasible and/or useful in the studied species and, if so, what features to look for. For example, are there matching phylogenetic patterns for whole regions of metabolism which link to the function of a transporter; or are single genes at key points in pathways a better indication of association of a transporter with the pathway. I, therefore, first investigated the characterised transporters which showed a variation in phylogenetic profile within the nematodes and/or platyhelminthes using phylogenetic profiling.

Characterised transporters: SLC25A21

SLC25A21 (the mitochondrial oxodicarboxylate carrier) acts to transport 2-oxoadipate into the mitochondrial matrix, in exchange for 2-oxoglutarate movement out of the matrix (Palmieri *et al.* 2001a; Fiermonte *et al.* 2001). Within the platyhelminthes, an orthologue of *SLC25A21* was predicted in the free-living platyhelminth *Schmidtea mediterranea*, but not in any of the studied liver flukes, blood flukes or tapeworms (*Figure 4.4*). Clustering of the orthologue dataset showed forty-eight additional IMPI positive training set genes with the same phylogenetic pattern, including two additional members of the mitochondrial carrier family (*SLC25A16/SLC25A43*) (*Appendix II – Table 1*). I analysed these genes (excluding the mitochondrial carriers) using the Reactome Pathway Knowledgebase (Fabregat *et al.* 2018) to identify overrepresented pathways. The top metabolic pathways are listed in *Table 4.2*.

Table 4.2 Top enriched Reactome metabolic pathways for IMPI genes only identified in the free-living platyhelminth *Schmidtea mediterranea*.

Reactome pathway	Count	<i>p</i> -value	Benjamini corrected <i>p</i> -value
Choline catabolism	4	2.10×10^{-6}	3.40×10^{-4}
Urea cycle	2	7.35×10^{-3}	0.19
Lysine catabolism	2	1.35×10^{-2}	0.19
Fatty acid metabolism	5	2.93×10^{-2}	0.19

One of the top enriched pathways was lysine catabolism (Reactome ID: R-HAS-71064). There are two routes of lysine catabolism in humans: the saccharopine pathway, located in the mitochondria (Higashino *et al.* 1965, 1967), and the pipecolate pathway, located in the peroxisomes and cytoplasm (Wanders *et al.* 1988, 1989). The pipecolate pathway is utilised mainly in the adult brain, whereas the saccharopine pathway dominates in other tissues (Chang 1976). The two genes of the lysine catabolic pathway identified by phylogenetic profiling were: *AADAT* and *ALDH7A1*. The proteins encoded by these genes function in the latter part of the saccharopine pathway, leading to the production of 2-oxoadipate – one of the main identified substrates of the carrier encoded by *SLC25A21* (Figure 4.9). So, it appears in this case that phylogenetic profiling provided useful information to predict the function of a mitochondrial carrier.

However, the saccharopine pathway is localised to the mitochondria, and so 2-oxoadipate is produced within the mitochondrial matrix. The main function of *SLC25A21* is not to transport 2-oxoadipate out of the matrix, but to bring 2-oxoadipate into the matrix in exchange for mitochondrial 2-oxoglutarate. Therefore, I expanded my phylogenetic analysis out into the parts of metabolism surrounding *SLC25A21*, in both the cytoplasm and the mitochondria (Figure 4.9), to investigate whether pathways producing 2-oxoadipate outside the mitochondria had the same phylogenetic profile.

Two further mitochondrial genes with the same phylogenetic pattern could be linked into the metabolism of 2-oxoadipate: *DHTKD1*, which catalyses the production of glutaryl-CoA from 2-oxoadipate, and *SUGCT*, which also catalyses the production of glutaryl-CoA but this time from glutarate. Using the MitoCore model of central metabolism (Smith *et al.* 2017) and the KEGG database of reactions (Kanehisa *et al.* 2016), I identified tryptophan breakdown in the cytoplasm as an additional site of 2-oxoadipate production, and therefore of interest to study phylogenetically. Six of the known genes of tryptophan catabolism also show the same phylogenetic profile as the carrier *SLC25A21*. Though the genes encoding the proteins of the last two reactions of the pathway, which lead to the production of 2-oxoadipate, are unknown, it can be assumed that this route of tryptophan metabolism as a whole has been lost in liver flukes, blood flukes, and tapeworms. This has led to the loss of the 2-oxoadipate carrier, as import of this metabolite into the mitochondria is no longer necessary.

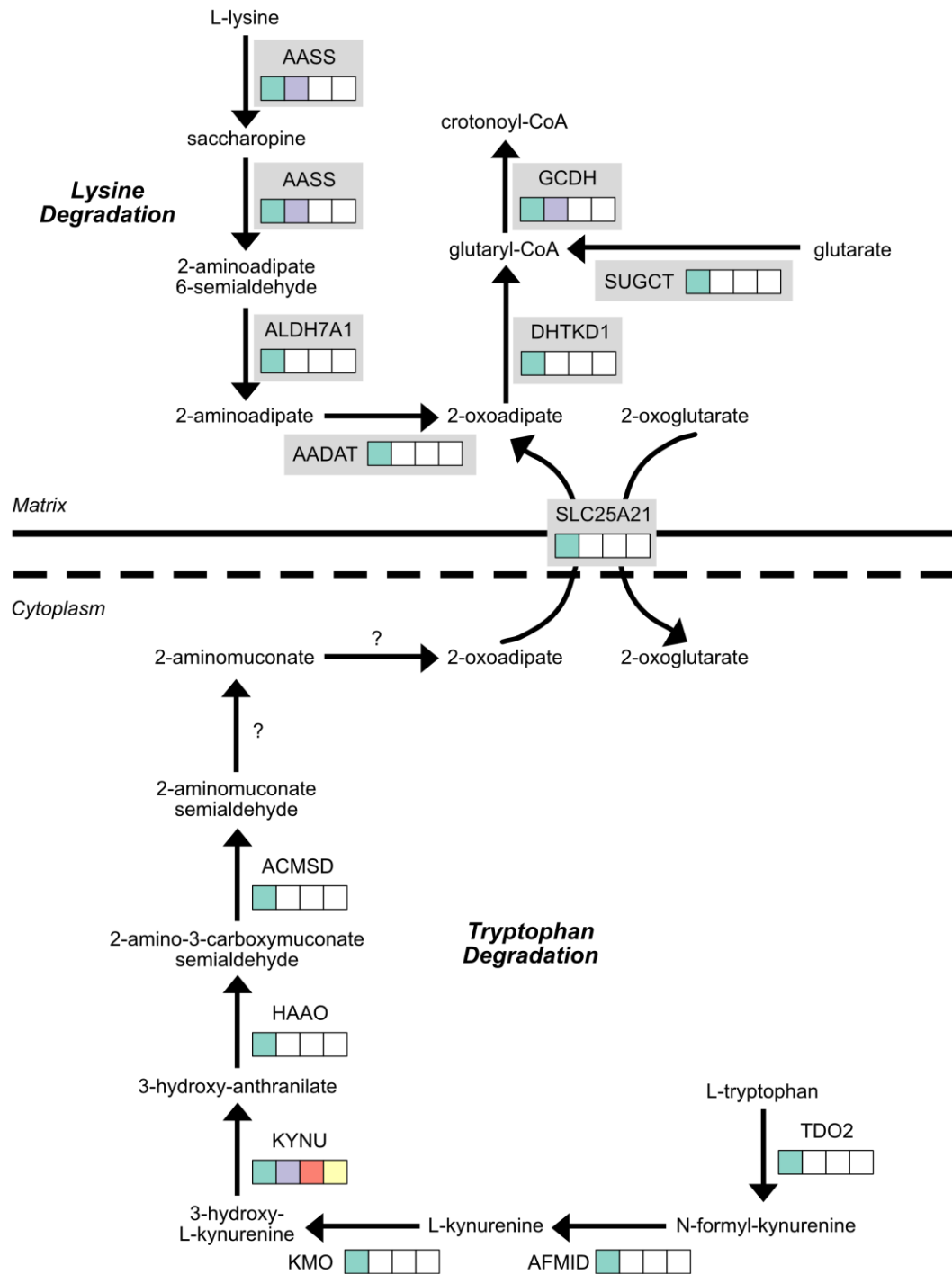


Figure 4.9 Phylogenetic profile of lysine and tryptophan catabolism in platyhelminthes, which matches the platyhelminth phylogenetic profile of the 2-oxoadipate carrier. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Grey boxes mark IMPI positive training genes.

However, the phylogenetic pattern of lysine and tryptophan catabolism does not appear to be linked to that of *SLC25A21* in the studied nematodes (*Figure 4.10*).

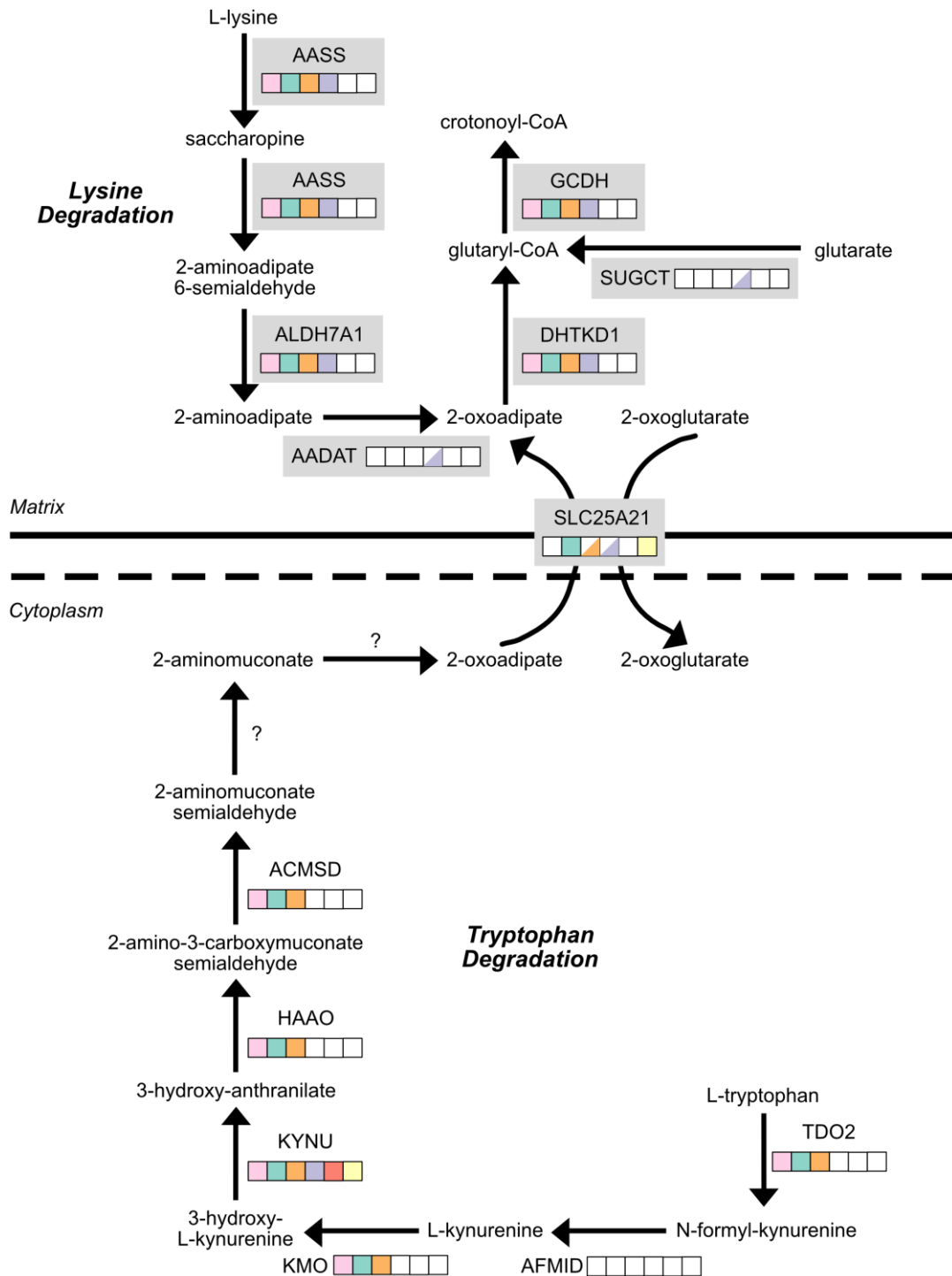


Figure 4.10 Phylogenetic profile of lysine and tryptophan catabolism in nematodes. Coloured boxes indicate that the gene is present in the following nematode species: pink = free-living *Diplogasterida*; green = free-living *Rhabditida*; orange = parasitic *Rhabditida*; purple = *Ascaridida*; red = *Filarioidea*; yellow = *Dorylaimia*. Half-filled boxes indicate this protein is missing in some of the species of that group. Grey boxes mark IMPI positive training genes.

SLC25A21 is identified in the *Dorylaimia*, but the genes of tryptophan and lysine catabolism are not; whilst the genes of tryptophan and lysine catabolism are identified in *Pristionchus pacificus*, but without identification of the *SLC25A21* transporter. The latter could be due to variability in the quality of protein annotation and orthologue prediction within this single species. However, the former is more problematic for use of phylogenetic profiling within the nematodes.

SLC25A21 provides a good study of what may be necessary to successfully implement phylogenetic profiling to identify substrates of carriers. A wider examination of metabolism around any possible transport substrate, including reactions located in the cytoplasm, may be useful in identifying the transported substrate.

Characterised transporters: SLC25A38

SLC25A38 has been genetically characterised as a glycine carrier important in haem metabolism (Fernández-Murray *et al.* 2016), though this has not been proven through transport assays. There has also been speculation in the literature that this transporter may carry 5-aminolevulinate produced from glycine out of the mitochondria (Guernsey *et al.* 2009). In my phylogenetic analysis of platyhelminthes, *SLC25A38* is identified in the free-living *S. mediterranea*, all liver flukes and tapeworms, but missing in the blood flukes (Figure 4.4). I identified thirteen genes showing the same pattern of phylogenetic loss in the blood flukes, including *SLC25A38* and an isoform of the ATP-Mg²⁺/P_i carrier *SLC25A24*. This number also includes some genes which are missing in one other studied platyhelminth, to allow some leeway for incomplete protein annotation or orthologue prediction.

The top seven pathways identified through Reactome analysis are all variations on mitochondria fatty acid β -oxidation (Table 4.3), which involve only the genes *ECHS1* (enoyl-CoA hydratase, short chain 1) and *HADH* (hydroxyacyl-CoA dehydrogenase). The number of counted genes in these identified pathways was three, due to the Reactome analysis associating two UniProt protein IDs with *HADH*. The gene *CPT2* (carnitine palmitoyltransferase 2) also shows the same phylogenetic pattern and is associated with fatty acid β -oxidation, producing acyl-CoAs from acyl-carnitines, which go on to enter the β -oxidation spiral (Ramsay *et al.* 2001). As fatty acid β -oxidation is not directly linked to

glycine or haem metabolism, the pathway enrichment analysis does not provide an appropriate prediction for the function of *SLC25A38*.

Table 4.3 Top enriched Reactome metabolic pathways for *IMPI* positive training genes, which are only missing from blood flukes of the studied platyhelminthes.

Reactome pathway	Count	<i>p</i> -value	Benjamini corrected <i>p</i> -value
β-oxidation of hexanoyl-CoA to butanoyl-CoA	3	1.29×10^{-6}	2.33×10^{-5}
β-oxidation of lauroyl-CoA to decanoyl-CoA-CoA	3	1.29×10^{-6}	2.33×10^{-5}
β-oxidation of octanoyl-CoA to hexanoyl-CoA	3	1.29×10^{-6}	2.33×10^{-5}
β-oxidation of decanoyl-CoA to octanoyl-CoA-CoA	3	2.05×10^{-6}	2.87×10^{-5}
Mitochondrial fatty acid β-oxidation of saturated fatty acids	3	3.22×10^{-5}	3.55×10^{-4}
β-oxidation of butanoyl-CoA to acetyl-CoA	2	2.33×10^{-4}	1.93×10^{-3}
Mitochondrial fatty acid β-oxidation	3	2.41×10^{-4}	1.93×10^{-3}

It may be that key individual genes with matching profiles can give clues to the function of a mitochondrial carrier. Therefore, I then went on to do a gene-by-gene analysis of the remaining eight genes with matching phylogenetic profiles in the platyhelminthes (*Appendix II – Table 2*). Three of these genes have characterised functions related to metabolic pathways. One was *CLYBL* – a malate/β-methylmalate synthase, which has been linked to the mitochondrial vitamin B12 pathway (Strittmatter *et al.* 2014). Vitamin B12 metabolism could be linked to haem metabolism in some way, as both molecules contain porphyrin rings. Vitamin B12 could also be linked to fatty acid β-oxidation, as a cofactor derived from vitamin B12 is necessary for the oxidation of propionyl-CoA, produced by β-oxidation of odd chain fatty acids. Whilst this adds further evidence to the idea that fatty acid β-oxidation is different in blood flukes than other platyhelminthes, no other mitochondrial gene known to be involved of Vitamin B12 metabolism shows the same phylogenetic pattern.

The second was *GLYCK* which catalyses the phosphorylation of *D*-glycerate in serine and fructose catabolism. Mutations in *GLYCK* had been identified in a patient who also had hyperglycinemia (Sass *et al.* 2010). It was hypothesised that this was caused by a blockage to glycine cleavage, through an uncharacterised system. However, more recent work has shown that this single patient had an additional mutation in the glycine cleavage enzyme *AMT*, rather

than *GLYCTK* malfunction influencing glycine levels (Swanson *et al.* 2017). Therefore, there was no clear link to glycine metabolism.

The third gene was *ABCB6* – an outer mitochondrial membrane protein involved in haem biosynthesis (Krishnamurthy *et al.* 2006). Assuming the genetic characterisation is correct and *SLC25A38* functions as a glycine transporter, I decided to look further at the metabolism around *ABCB6*, as glycine transport is also linked to haem metabolism. Though the two transporters both show the same phylogenetic pattern in platyhelminthes, the haem catabolism pathway that the transporters are linked to do not (*Figure 4.11*).

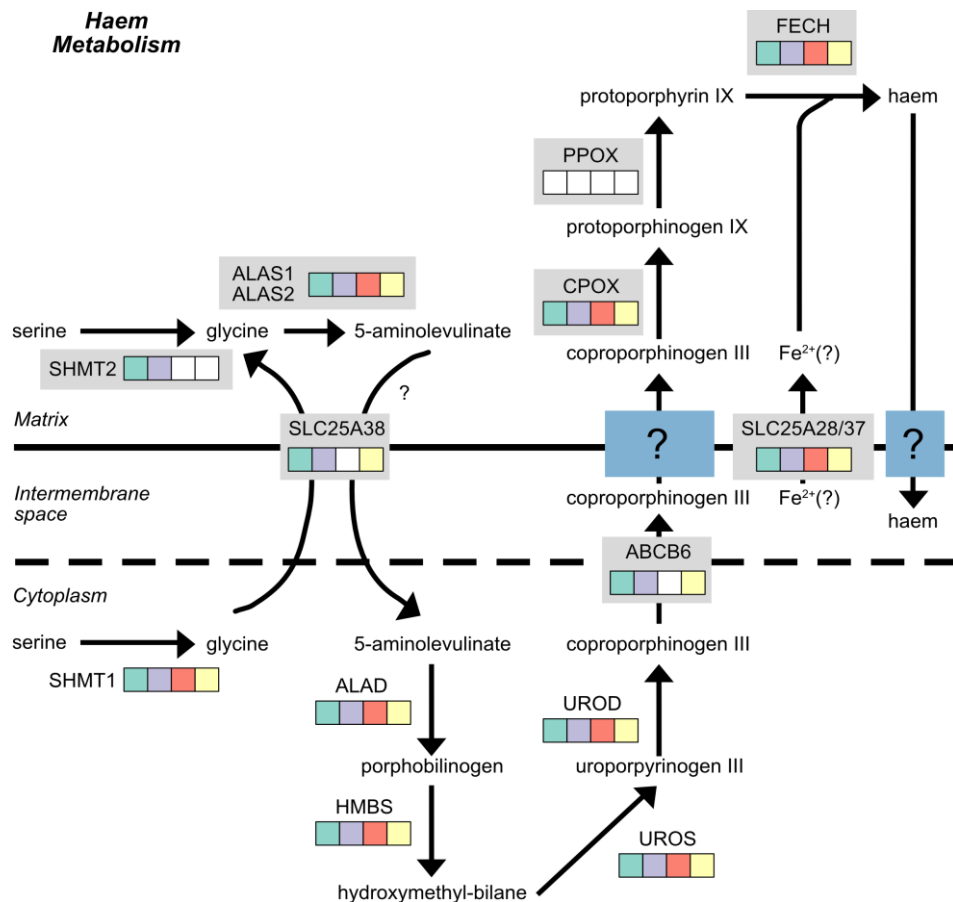


Figure 4.11 Phylogenetic profile of haem metabolism in platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Grey boxes mark IMPI positive training genes.

Most of the haem biosynthesis enzymes in humans are conserved across all the studied platyhelminthes, confirming haem metabolism as a key pathway. Only the known transporters are lost in blood flukes, though not all transport steps have identified genes or are fully characterised.

Haem synthesis genes have been lost in nematodes (Rao *et al.* 2005). This is confirmed in my analysis of nematode species (*Figure 4.12*).

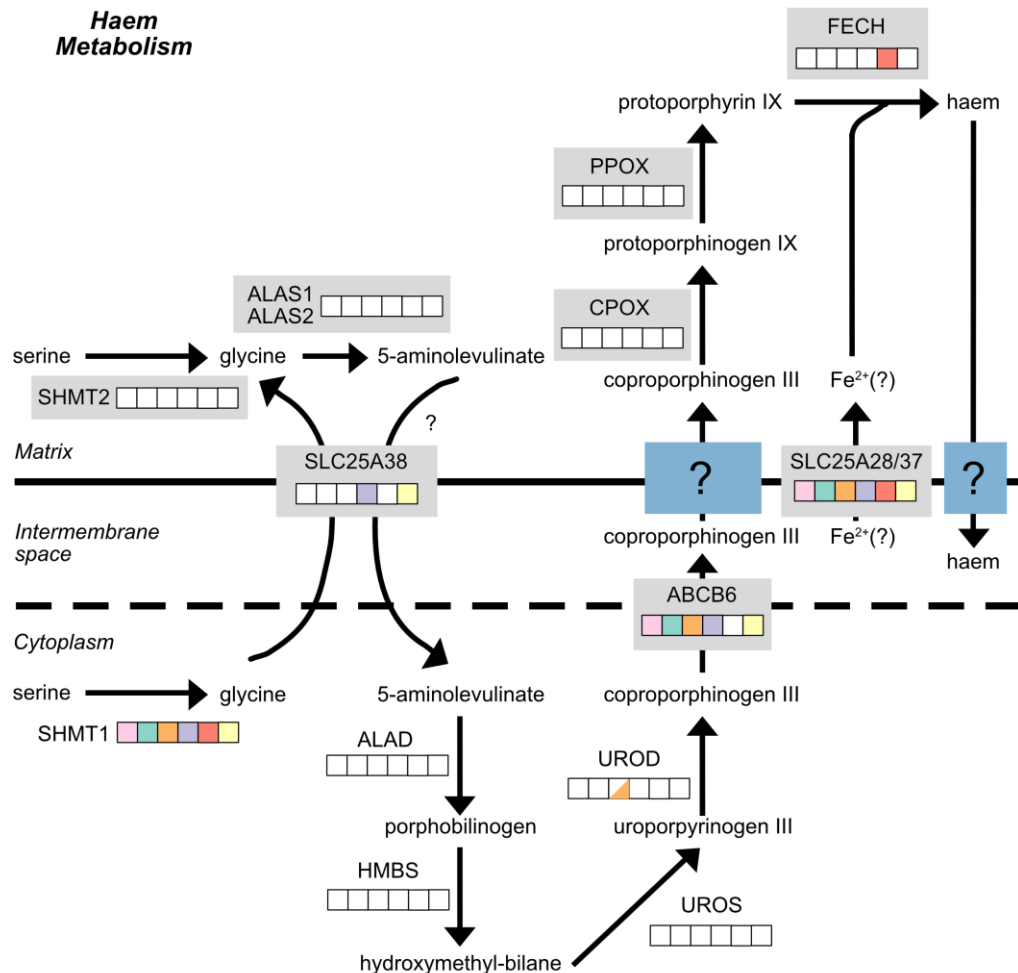


Figure 4.12 Phylogenetic profile of haem metabolism in nematodes. Coloured boxes indicate that the gene is present in the following nematode species: pink = free-living *Diplogasterida*; green = free-living *Rhabditida*; orange = parasitic *Rhabditida*; purple = *Ascaridida*; red = *Filarioidea*; yellow = *Dorylaimia*. Half-filled boxes indicate this protein is missing in some of the species of that group. Grey boxes mark IMPI positive training genes.

However, an orthologue of *ABCB6* has been identified in *C. elegans* (*hmt-1*) and is associated with heavy metal (including iron) detoxification (Schwartz *et al.* 2010). Dependent on the localisation of the haem insertion into proteins, haem may still have to enter the mitochondria in these species (Kim *et al.* 2012). Therefore, there may be a haem transporter in these nematodes, though it may function mainly to import rather than export haem.

The loss of only the transport steps of haem metabolism in blood flukes may suggest alternate localisation of some haem metabolic steps in these species – if steps of haem metabolism are moved from the mitochondria, there would be no need for the transport steps. Therefore, I investigated the predicted mitochondrial targeting scores of the mitochondrial localised parts of haem metabolism across the platyhelminthes and several model organisms, using four different mitochondrial targeting sequence prediction programs (*Table 4.4*).

In this case, the predicted mitochondrial targeting scores are not very consistent either within species or between species. Between species differences may be exaggerated by less accurate protein annotation, as mitochondrial targeting prediction programs require an accurate sequence at the *N*-terminal of the protein. However, it does not appear that there has been a loss of mitochondrial localisation of *ALAS*, *CPOX*, and *FECH* specifically in the blood flukes, as these species have some of the highest targeting scores in the studied platyhelminth species. Therefore, differences in localisation of parts of haem metabolism does not seem to explain the loss of the *SLC25A38* carrier in the blood flukes.

In the studied nematodes, only a single IMPI positive training gene matches the phylogenetic profile of *SLC25A38* – not identified in the Diplogasterida, most Rhabditida and the Filarioidea. That gene (*SLC44A1*) is a choline transporter, which is partially localised to the mitochondrial membrane, allowing transport of choline in and out of the mitochondria (Michel & Bakovic 2009). Choline is catabolised to glycine, mainly in the mitochondria, providing a possible link to the function of *SLC25A38*.

Table 4.4 Predicted mitochondrial targeting sequence scores for the predicted orthologues of the mitochondrial enzymes of haem metabolism in the studied platyhelminthes and humans, mouse and zebrafish. iP = iPSORT, MP = MitoProt II, TP = TargetP, MF = MitoFates. Yellow boxes are scores over 0.75.

Species	ALASI						ALAS2						CPOX						FECH					
	iP	MP	TP	MF	iP	MP	TP	MF	iP	MP	TP	MF	iP	MP	TP	MF	iP	MP	TP	MF	TP	MF	TP	MF
<i>Homo sapiens</i>	1	0.97	0.72	0.98	0	0.84	0.35	0.99	0	0.60	0.46	1.00	0	0.67	0.88	1.00								
<i>Mus musculus</i>	1	0.98	0.70	1.00	1	0.84	0.46	0.87	1	0.32	0.75	0.33	0	0.40	0.79	1.00								
<i>Danio rerio</i>	1	0.95	0.61	0.1	-	-	-	-	1	1.00	0.91	0.75	1	0.88	0.86	0.89								
<i>Schmidtea mediterranea</i>	0	0.55	0.50	0.47	-	-	-	-	0	0.41	0.07	0.02	1	0.65	0.31	0.98								
<i>Clonorchis sinensis</i>	0	0.44	0.25	0.15	-	-	-	-	0	0.03	0.08	0.00	0	0.08	0.16	0.00								
<i>Opisthorchis viverrini</i>	0	0.27	0.27	0.18	-	-	-	-	0	0.01	0.11	0.00	0	0.16	0.08	0.00								
<i>Schistosoma mansoni</i>	1	0.89	0.68	0.15	-	-	-	-	1	0.56	0.83	0.43	0	0.75	0.85	0.48								
<i>Schistosoma japonicum</i>	1	0.85	0.50	0.13	-	-	-	-	0	0.23	0.55	0.03	0	0.93	0.17	0.09								
<i>Schistosoma haematobium</i>	1	0.93	0.68	0.14	-	-	-	-	0	0.42	0.60	0.13	1	0.66	0.76	0.43								
<i>Echinococcus granulosus</i>	1	0.88	0.65	0.39	-	-	-	-	0	0.26	0.19	0.04	0	0.94	0.79	0.74								
<i>Echinococcus multilocularis</i>	1	0.89	0.69	0.41	-	-	-	-	0	0.31	0.18	0.04	1	0.76	0.61	0.45								
<i>Hymenolepis microstoma</i>	0	0.34	0.23	0.21	-	-	-	-	0	0.03	0.06	0.01	0	0.18	0.12	0.08								
<i>Taenia solium</i>	0	0.18	0.12	0.08	-	-	-	-	0	0.64	0.46	0.35	0	0.3	0.09	0.00								

However, none of the remaining genes of choline catabolism show the same phylogenetic profile in nematodes (*Figure 4.13*), casting doubt on this association.

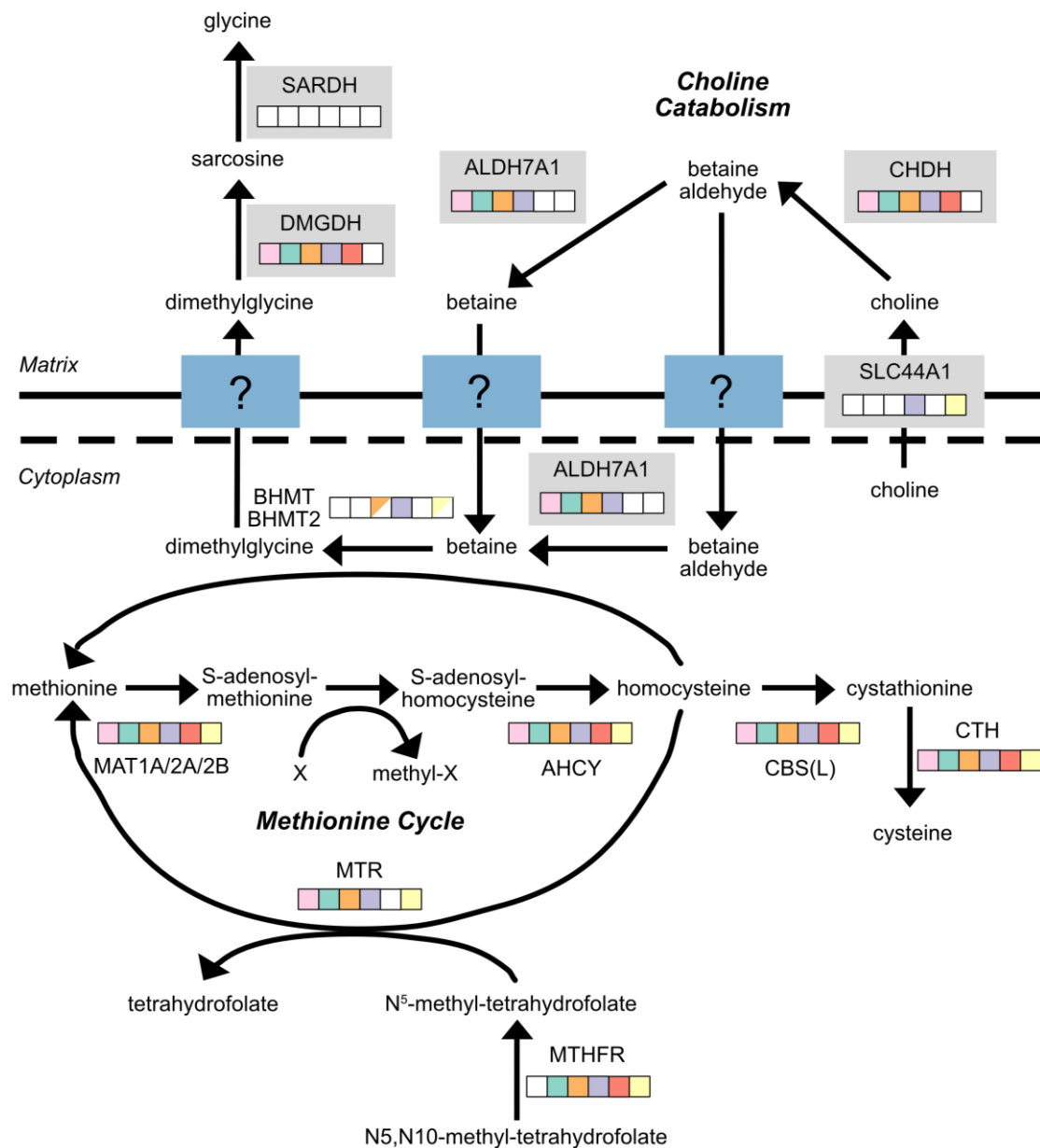


Figure 4.13 Phylogenetic profile of choline catabolism and the methionine cycle in studied nematodes. Coloured boxes indicate that the gene is present in the following nematode species: pink = free-living Diplogasterida; green = free-living Rhabditida; orange = parasitic Rhabditida; purple = Ascaridida; red = Filarioidea; yellow = Dorylaimia. Half-filled boxes indicate this protein is missing in some of the species of that group. Grey boxes mark IMPI positive training genes.

In summary, phylogenetic profiling in both nematodes and platyhelminthes does not definitively link the *SLC25A38* transporter with its genetically characterised substrate (glycine), though there are small clues that may point to this being the case. Large parts of the haem synthesis pathway in humans are identified in all of the studied platyhelminthes, including the blood flukes; although an additional mitochondrial transporter linked to haem synthesis (*ABCB6*) does show the same phylogenetic profile as *SLC25A38* in platyhelminthes. It does not appear that there is a change in localisation of the mitochondrial parts of haem metabolism in blood flukes, which could have been linked to loss of the carrier.

Characterised transporters: SLC25A12/13

SLC25A12 and *SLC25A13* are transporters which exchange mitochondrial aspartate for cytoplasmic glutamate (Palmieri *et al.* 2001b). Together with the oxoglutarate carrier (*SLC25A11*), these carriers facilitate the functioning of the malate-aspartate shuttle, supporting the import of reducing equivalents (NADH) into the mitochondrial matrix (Amoedo *et al.* 2016). A predicted orthologue of *SLC25A12/13* is identified in the free-living platyhelminth *S. mediterranea*, liver flukes and most blood flukes, but only in *T. solium* of the four studied tapeworms (*Figure 4.4*).

No IMPI positive training genes match the phylogenetic pattern of *SLC25A12/13* in platyhelminthes, and no other genes are only identified in *T. solium* of the four different tapeworm species studied, regardless of the rest of the phylogenetic pattern. The genes encoding members of the malate-aspartate shuttle, including the oxoglutarate carrier, are completely or almost completely present in all studied platyhelminthes (*Figure 4.14*).

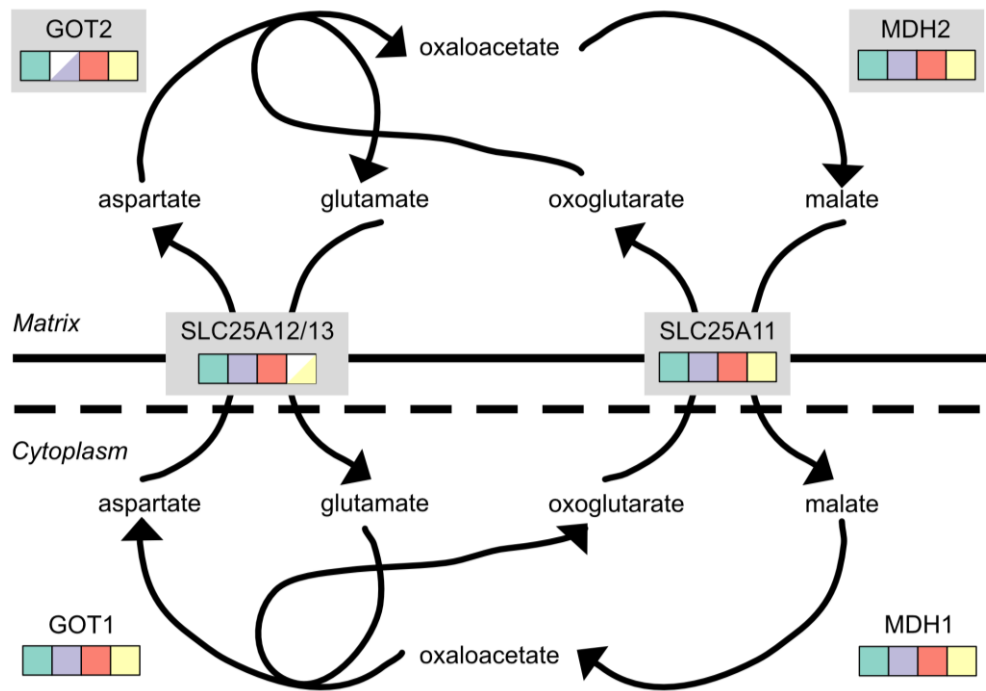


Figure 4.14 Phylogenetic profile of the malate-aspartate shuttle in platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Half-filled boxes indicate this protein is missing in some of the species of that group. Grey boxes mark IMPI positive training genes.

The malate-aspartate shuttle is linked to the tricarboxylic acid (TCA) cycle, through the production of malate from oxaloacetate by *MDH2* (malate dehydrogenase 2). In the TCA cycle, malate dehydrogenase catalyses the reaction in the opposite direction, producing oxaloacetate from malate. Genes encoding all other enzymes of the TCA cycle are identified in all the studied platyhelminth species, except fumarase (*FH*) which is not identified in any tapeworm. However, there are two known classes of fumarase – type I and type II. The fumarase identified in humans and the rest of the platyhelminth species is a type II fumarase (Estévez *et al.* 2002). All studied tapeworm species instead have a type I fumarase which catalyses the same reaction, but which forms iron-dependent dimers (Estévez *et al.* 2002). Therefore, the TCA cycle looks to be complete in all studied platyhelminthes and is not linked phylogenetically to the aspartate/glutamate carrier.

The inconsistency of the phylogenetic pattern in the tapeworms raises the possibility that the prediction of an orthologue in *T. solium* may actually be caused by a contaminant (i.e. this transporter group is actually lost from all tapeworms).

In the studied nematode species, a predicted orthologue of *SLC25A12/13* is missing in the Filarioidea and *Pristionchus pacificus* (Figure 4.5). The loss in all three studied Filariod nematodes is particularly interesting for its consistency within that taxon; whilst *P. pacificus* is the only diplogasterid included in the analysis, so the loss is not supported by analysis of related species. Five IMPI positive training set genes exactly match the phylogenetic pattern of the aspartate/glutamate carriers within the nematodes: *PAICS*, *MTHFS*, *MMADHC*, *CHCHD4* and *MCEE*. *PAICS* (involved in purine synthesis) consumes aspartate, though this protein is not solely localised to the mitochondria (French *et al.* 2016). Two of these genes (*MMADHC* and *MCEE*) are related to vitamin B12 metabolism. This may be of some interest as vitamin B12 dependent enzymes are utilised in some methylation reactions dependent on tetrahydrofolate; and tetrahydrofolate itself is synthesised using glutamate. An additional fourteen genes are lost in the Filarioidea but identified in *P. pacificus*, with another nine genes with a similar pattern but missing between one and three orthologues from other species. I took all these genes forward for further analysis.

Reactome pathway enrichment analysis of these genes identifies the complete loss of the pathway ‘propionyl-CoA catabolism’, which produces succinyl-CoA from propionyl-CoA and is dependent on adenosylcobalamin (a derivative of vitamin B12) as a cofactor. Most of the remaining identified pathways are related to vitamin and cofactor metabolism, with the strongest *p*-value for the pathway ‘defects in vitamin and cofactor metabolism’ (Benjamini corrected *p*-value = 2.34×10^{-11}). While this is interesting for potential future profiling, this does not link the phylogenetic pattern to the function of *SLC25A12/13*.

However, one of the genes with a matching or close phylogenetic pattern to *SLC25A12/13* in the nematodes is *GOT2* (mitochondrial aspartate transaminase), which is a key part of the malate-aspartate shuttle. I investigated the other genes of the malate-aspartate shuttle in the nematodes (Figure 4.15), but there was no loss of any of the cytoplasmic genes of the malate-aspartate shuttle.

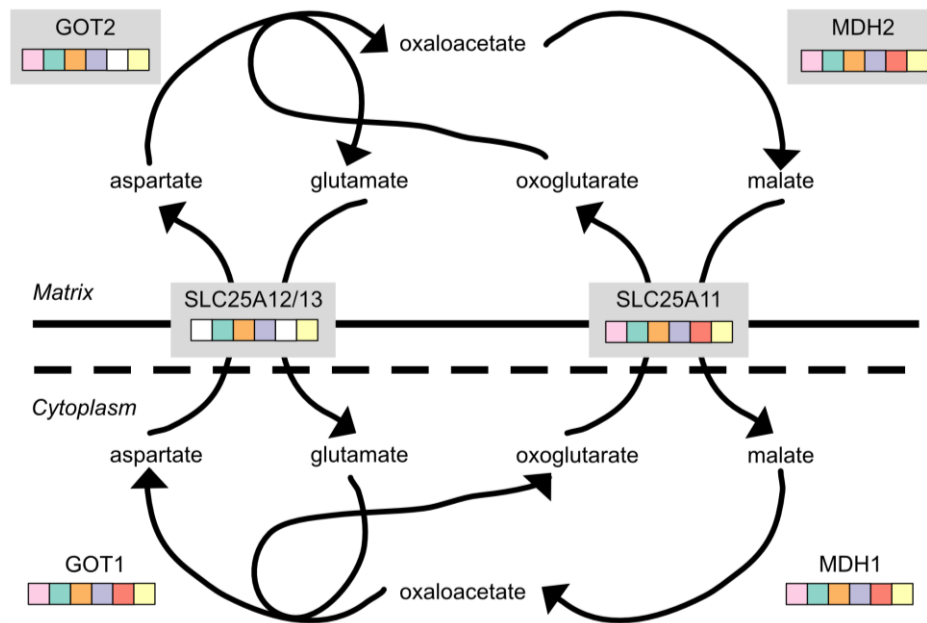


Figure 4.15 Phylogenetic profile of the malate-aspartate shuttle in nematodes. Coloured boxes indicate that the gene is present in the following nematode species: pink = free-living *Diplogasterida*; green = free-living *Rhabditida*; orange = parasitic *Rhabditida*; purple = *Ascaridida*; red = *Filarioidea*; yellow = *Dorylaimia*. Grey boxes mark IMPI positive training genes.

Though this may indicate a small potential link between gene phylogenetic profiles and the function of the aspartate/glutamate carrier, it is only one gene from twenty-nine with similar patterns. It seems unlikely that this would provide enough evidence to justify wet laboratory testing of the carrier if the function was still unknown, even if I had managed to pick out the correct gene. Therefore, the phylogenetic pattern analysis of the aspartate/glutamate carrier provides some important lessons for the application of this technique. Links to the function of the gene may be very small (in this case one gene) and difficult to detect, and it will be important to use other corroborating evidence to justify wet laboratory work.

Characterised transporters: SLC25A2/15

SLC25A2 and *SLC25A15* are both characterised ornithine transporters, which transport cytoplasmic ornithine into the mitochondrial matrix, in exchange for mitochondrial citrulline (Camacho *et al.* 1999, 2003). A predicted orthologue of the ornithine transporters is identified in most studied nematode species with no consistent pattern of loss (Figure 4.5). However, in

the platyhelminthes, an ornithine transporter is not predicted in either of the two *Echinococcus* species, despite identification in the other two tapeworm species (Figure 4.4).

The key function of ornithine in human mitochondria is the detoxification of ammonia via the urea cycle, where ornithine is combined with carbamoyl-phosphate to form citrulline, via the protein encoded by the gene *OTC* (ornithine carbamoyltransferase). The genes of the urea cycle show a very inconsistent phylogenetic pattern within the platyhelminthes (Figure 4.16), with different parts lost throughout the studied species. The entire pathway, apart from cytoplasmic arginase (*ARG1*) is lost in *Echinococcus* species, but also in the other studied tapeworms.

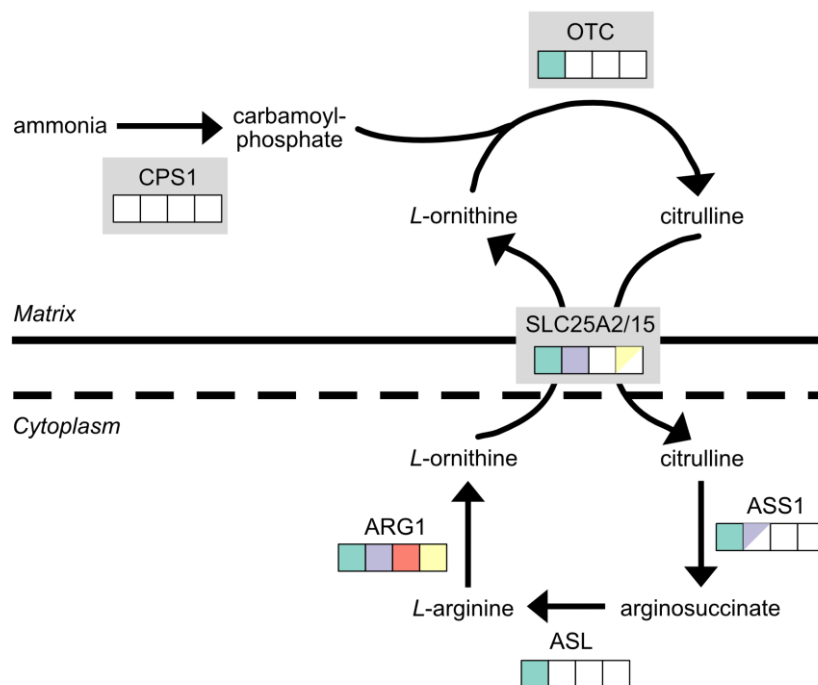


Figure 4.16 Phylogenetic profile of the urea cycle in platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Half-filled boxes indicate this protein is missing in some of the species of that group. Grey boxes mark IMPI positive training genes.

No IMPI positive training genes exactly match the phylogenetic pattern of the ornithine carriers in the platyhelminthes. However, five genes are identified in only one of *T. solium* and *H. microstoma* (the two non-*Echinococcus* tapeworm species), listed in Table 4.5.

Table 4.5 IMPI positive training genes with similar phylogenetic patterns to the ornithine carriers in platyhelminthes.

Ensembl ID	Gene	Function
ENSG00000131473	<i>ACLY</i>	Synthesis of acetyl-CoA
ENSG00000159423	<i>ALDH4A1</i>	Proline degradation
ENSG00000162129	<i>CLPB</i>	Peptidase
ENSG00000100033	<i>PRODH</i>	Proline degradation
ENSG00000115840	<i>SLC25A12</i>	Aspartate/glutamate exchanger

Two of these genes (*ALDH4A1* and *PRODH*) are involved in proline degradation to glutamate via *L*-pyrroline-5-carboxylate (P5C), but I could find no described link between *SLC25A2/15* and proline transport in the literature. P5C can also be produced from ornithine via the action of ornithine aminotransferase (*OAT*), producing *L*-glutamate 5-semialdehyde. This can then be non-enzymatically converted to P5C (Mezl & Knox 1976). Predicted orthologues of *OAT*, however, are identified in both species of *Echinococcus* studied – i.e. this step has not been lost. In rat liver tissue, *OAT* can run in reverse, producing ornithine from *L*-glutamate 5-semialdehyde, when there are low concentrations of ornithine (Strecker 1965). This may be the case in *Echinococcus* mitochondria, if the loss of the ornithine transporter leads to a lack of transport of ornithine into the matrix, explaining the presence of *OAT* in *Echinococcus*.

In summary, small parts of proline metabolism linked to ornithine show a similar phylogenetic pattern to the ornithine transporters in the platyhelminthes. However, this link would not be obvious if the ornithine transporter was not already functionally characterised.

Characterised transporters: SLC25A10

The dicarboxylate carrier (*SLC25A10*) transports dicarboxylates, including malate and succinate, out of the mitochondrial matrix in exchange for inorganic phosphate (P_i) (Palmieri *et al.* 1996; Fiermonte *et al.* 1998). Although orthologues of this carrier are almost ubiquitously predicted in the studied platyhelminthes (*Figure 4.4*), predicted orthologues of this gene are not predicted in any of the studied species of *Caenorhabditis* or in *P. pacificus* (*Figure 4.5*). Three additional genes have similar phylogenetic patterns to this transporter within the nematodes, but none of these have clear metabolic functions (*Table 4.6*). There are,

therefore, no clues in the phylogenetic profiling of mitochondrial genes in the nematodes and platyhelminthes as to the function of *SLC25A10*.

Table 4.6 *IMPI positive training genes with similar phylogenetic patterns to the dicarboxylate carrier in nematodes.*

Ensembl ID	Gene	Function
ENSG00000114026	<i>OGG1</i>	DNA repair
ENSG00000164306	<i>PRIMPOL</i>	DNA primase-polymerase
ENSG00000156990	<i>RPUSD3</i>	Mitochondrial rRNA modification

Characterised transporters: SLC25A1 and SLC25A29

SLC25A1 encodes the mitochondrial citrate carrier, which exchanges mitochondrial citrate (or other tricarboxylates) for cytoplasmic malate (Bisaccia *et al.* 1989). This carrier is identified in all studied platyhelminthes but is missing from nematode members of the Dorylaimia (including both *Trichinella* and *Trichuris* species). Thirty IMPI positive training set genes have a similar phylogenetic pattern in nematodes (*Appendix II – Table 3*), but no metabolic pathways are significantly overrepresented in these genes in a Reactome pathway analysis – the highest ranked pathway is ‘mitochondrial protein import’ (Benjamini corrected p -value = 0.045).

The citrate exported from the mitochondrial matrix by *SLC25A1* can be utilised in the cytoplasm to feed acetyl-CoA to fatty acid and sterol synthesis, via the action of ATP citrate lyase (*ACLY*). Orthologues of *ACLY* are identified in all studied nematode species. There is a single IMPI positive training gene with a matching phylogenetic pattern linked to fatty acid biosynthesis – malonyl-CoA decarboxylase (*MLYCD*) – which catalyses the breakdown of malonyl-CoA to acetyl-CoA (Sacksteder *et al.* 1999). However, in general, there is no obvious link between the genes matching the phylogenetic pattern of *SLC25A1* and its function.

SLC25A29 (the basic amino acid transporter) also shows a similar phylogenetic pattern within the nematodes, although an orthologue is identified in a single member of the Dorylaimia (*T. spiralis*). Orthologues are not predicted in any studied platyhelminth species. There are no

IMPI 2017 positive training genes that exactly match this phylogenetic pattern. Additionally, none of the genes lost in all the *Dorylaimia* have a clear association with basic amino acid metabolism (*Appendix II – Table 3*). The identification of an orthologue in *T. spiralis* may suggest that orthologues in other very closely related species of *Trichinella* may have been missed due to incomplete protein annotation, or that the orthologous protein in *T. spiralis* is a contaminant.

In summary, phylogenetic profiling within the nematodes or platyhelminthes does not predict the function of either *SLC25A1* or *SLC25A29*.

Lessons from the characterised transporters

These investigations into the characterised transporters gave good information about the potential success rate and limitations of phylogenetic profiling to predict the function of mitochondrial carriers. The identified patterns of metabolic loss do not seem to be consistent between the platyhelminthes and the nematodes, presumably due to different modifications of metabolism in comparison to known human metabolism. The patterns of gene loss within certain types of platyhelminthes seem more consistent, and it is only within the platyhelminthes that it was possible to link the loss of metabolism to transporter function (*SLC25A21*). In other cases, single genes could be linked to carrier function via their phylogenetic pattern, but extensive extra work would have been needed to support a case for testing any predicted transport substrate from such a small association. Cases like these would be harder to identify for uncharacterised transporters. In other cases, there was no apparent link between carrier function and phylogenetic patterns of the metabolism linked to the transport substrate, suggesting this approach will not solve all cases.

Uncharacterised transporters

There are four groups of uncharacterised transporters which showed variation in the phylogenetic pattern of presence and absence in the platyhelminthes and/or the nematodes: *SLC25A14/30*, *SLC25A43*, *SLC25A44* and *SLC25A45/47/48* (Figures 4.4 and 4.5). I applied the knowledge gained from investigating the characterised transporters to inform my study of these uncharacterised mitochondrial carrier family transporters.

I also used additional information from the literature to provide more context to any potential transport substrate predicted by the phylogenetic analysis. This included looking at a symmetry/binding site analysis which predicted the potential type of transport substrate for a variety of carriers (Robinson *et al.* 2008). The binding site analysis looks at the predicted substrate binding sites across carrier groups and uses characterised transporters and chemical knowledge to predict the likely features of a transported substrate (Robinson *et al.* 2008). I also carried out a literature search for any other information on the carriers and their potential function.

Table 4.7 summarises the results of this analysis, before a more extensive case-by-case discussion of the four studied groups of uncharacterised transporters.

.

Table 4.7 Summary of the substrate predictions for four groups of uncharacterised mitochondrial carriers, with associated evidence.

Carrier(s)	Phylogenetic pattern		Additional substrate information		Potential substrate(s)
	Platyhelminthes	Nematodes	Symmetry analysis	Literature search	
SLC25A14/30	Present in <i>S. mediterranea</i> and the liver flukes. Missing in the blood flukes and tapeworms	None	Small keto acid	Linked to ketone body metabolism	Ketone bodies, branched chain α -keto acids, branched chain amino acids
SLC25A43	Present only in <i>S. mediterranea</i>	Present only in the Dorylaimia	Importer of a substrate with a nucleotide structure	Poorly covered in the literature	None
SLC25A44	Present in <i>S. mediterranea</i> , liver flukes and blood flukes. Missing from tapeworms	All	Substrate with a charged group and a hydrophobic or uncharged polar side chain	Poorly covered in the literature	Cyclic pyranopterin monophosphate (cPMP)
SLC25A45/47/48	Missing from tapeworms and possibly <i>Schistosoma japonicum</i>	Missing from the Filariodea and Dorylaimia	None	Poorly covered in the literature	None

Uncharacterised transporters: SLC25A14/30

SLC25A14 and *SLC25A30* (also known as *UCP5/BMCP1* and *KMCP1*) are two uncharacterised members of the mitochondrial carrier family. They group closely together in the human carrier phylogenetic tree (*Figure 4.3*), so I considered them together in the phylogenetic profiling analysis. The studied nematodes have no predicted orthologues of these genes (*Figure 4.5*). In the studied platyhelminthes, orthologues are predicted in the free-living species *S. mediterranea* and the liver flukes, but not in the blood flukes or tapeworms (*Figure 4.4*). Therefore, this analysis was concentrated on the platyhelminthes.

Thirty-five genes (*Appendix II – Table 4*) from the IMPI positive training set have phylogenetic profiles matching *SLC25A14/30* within the platyhelminthes. An additional nine genes were also included in further analysis, whose phylogenetic profile differed from that of *SLC25A14/30* by one species, to allow for variation in the quality of the protein annotation and orthologue identification processes. Reactome pathway enrichment analysis identified several pathways that are significantly enriched within these genes. The top four metabolic and non-overlapping pathways are listed in *Table 4.8*.

Table 4.8 Enriched Reactome metabolic pathways for genes matching the phylogenetic pattern of *SLC25A14/30*, within the platyhelminth species.

Reactome pathway	Count	<i>p</i> -value	Benjamini corrected <i>p</i> -value
Branched-chain amino acid catabolism	13	1.11×10^{-16}	9.99×10^{-15}
Mitochondrial fatty acid β -oxidation	8	3.12×10^{-9}	6.86×10^{-8}
Ketone body metabolism	4	3.03×10^{-6}	3.03×10^{-5}
Cobalamin (Cbl, vitamin B12) transport and metabolism	3	6.80×10^{-4}	2.72×10^{-3}

One of the significantly enriched pathways is ‘mitochondrial fatty acid β -oxidation’, which is the process by which fatty acids are broken down to acetyl-CoA (which enters the tricarboxylic acid cycle) and the reducing equivalents NADH and FADH₂ (Houten *et al.* 2016). Large parts of fatty acid β -oxidation have been lost in blood flukes and tapeworms, including the electron acceptor electron transfer flavoprotein (*Figure 4.17*). However, the transport system feeding fatty acids into the mitochondrial fatty acid β -oxidation spiral via

carnitine is well characterised (involving carnitine palmitoyltransferases *CPT1* and *CPT2* and the mitochondrial carrier *SLC25A20*) with no obvious additional transporters left to identify.

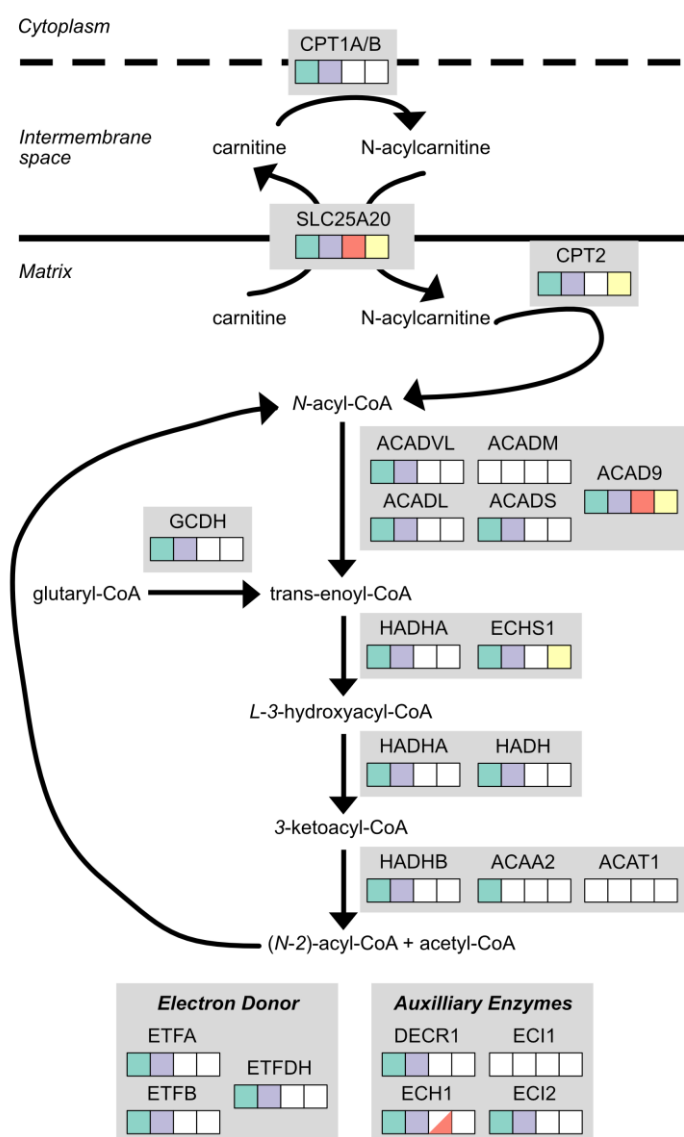


Figure 4.17 Phylogenetic profile of mitochondrial fatty acid β -oxidation in platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Grey boxes mark IMPI positive training genes. Half-filled boxes = missing in some species of that group.

The overrepresented metabolic pathway with the highest number of identified genes and the lowest corrected p -value is ‘branched-chain amino acid catabolism’ – the breakdown of valine, leucine and isoleucine. Most genes encoding parts of this pathway seem to have been lost in blood flukes and tapeworms (Figure 4.18), including genes encoding regulatory proteins, such as *BCKDK* (branched chain ketoacid dehydrogenase kinase) and *PPMIK*

(protein phosphatase 1K). The vitamin B12 metabolic pathway, also identified in the Reactome analysis, links into this pathway through the vitamin B12 dependent enzyme *MUT* (methylmalonyl-CoA mutase; *Figure 4.18*).

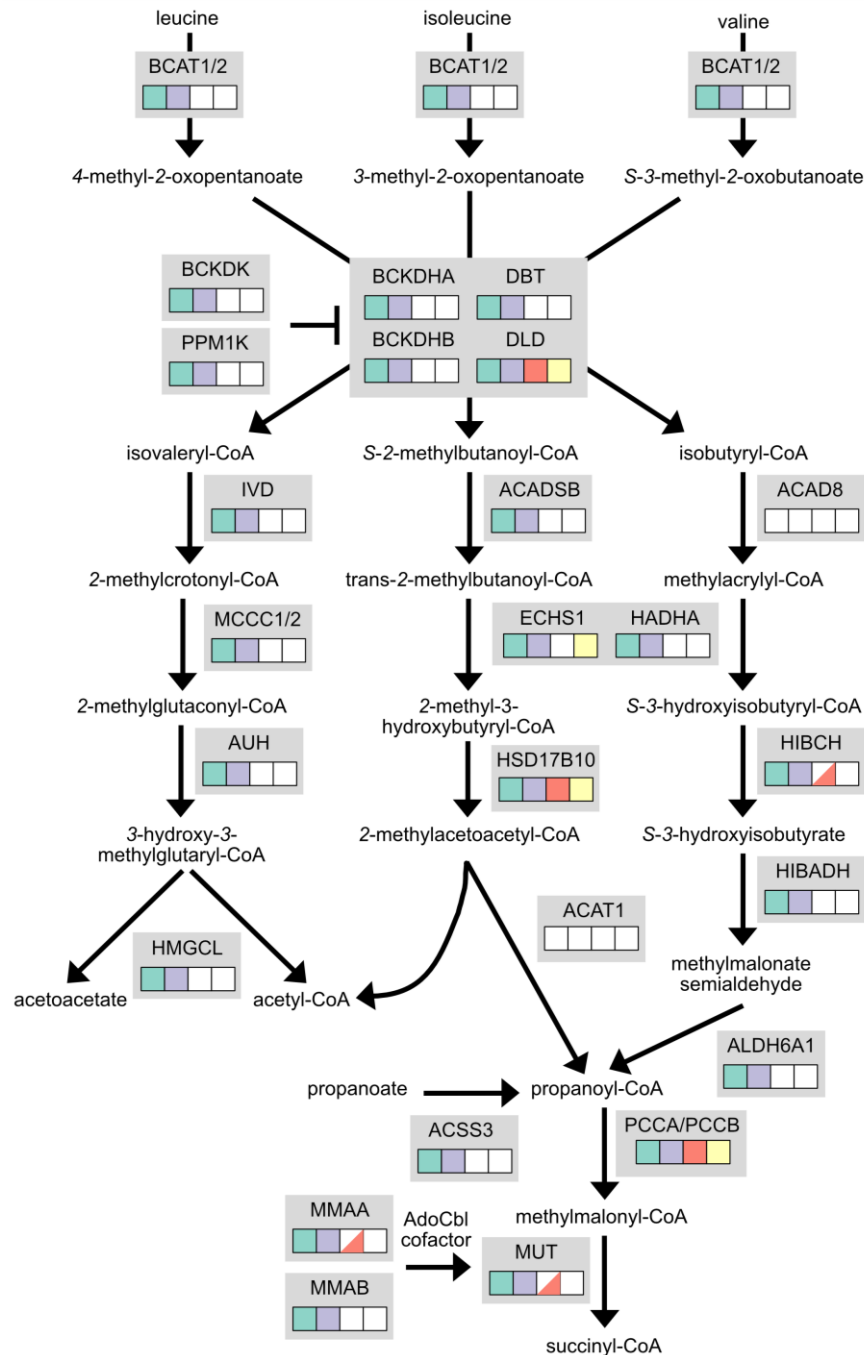


Figure 4.18 Phylogenetic profile of branched chain amino acid degradation in platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Grey boxes mark IMPI positive training genes. AdoCbl = adenosylcobalamin (vitamin B12 derivative). Half-filled boxes indicate this protein is missing in some of the species of that group.

In humans, this pathway can be localised entirely to the mitochondria. Therefore, any possible transport steps should be at the beginning and/or end of this pathway. The beginning of the pathway is the three amino acids (valine, leucine and isoleucine). Branched chain amino acid transaminase catalyses the first step of this pathway for all three amino acids, producing a related branched-chain keto acid. In humans, there are two genes encoding this enzyme: *BCAT1* (localised to the cytoplasm) and *BCAT2* (localised to the mitochondria). If this step is carried out in the cytoplasm, there is the potential for import of the keto acid for completion of the catabolic process within the mitochondria. Otherwise, the amino acids themselves must enter the mitochondria. Therefore, the beginning of the pathway provides two possible groups of potential substrates for a transporter – the three branched-chain amino acids and their corresponding branched-chain keto acids.

The end of the branched-chain amino acid catabolic pathway is different for the three amino acids involved. Valine and isoleucine can eventually be converted to succinyl-CoA, through different but complementary paths, which can enter the tricarboxylic acid cycle. The catabolism of leucine, however, leads to the production of acetyl-CoA and acetoacetate. This links into the third enriched Reactome pathway – ‘ketone body metabolism’ (*Figure 4.19*).

Acetoacetate is one of three ketone bodies in humans, the others being acetone and 3-hydroxybutanoate. Ketone bodies are produced by the liver in humans, mainly in the mitochondrial matrix, and used as an energy source by other tissues, particularly during periods of fasting (McGarry *et al.* 1970). To get to other tissues, or to be used in other processes such as sterol synthesis, ketone bodies must be transported out of the mitochondrial matrix. Therefore, a third possible group of transport substrates for *SLC25A14/30* identified from phylogenetic profiling are the ketone bodies.

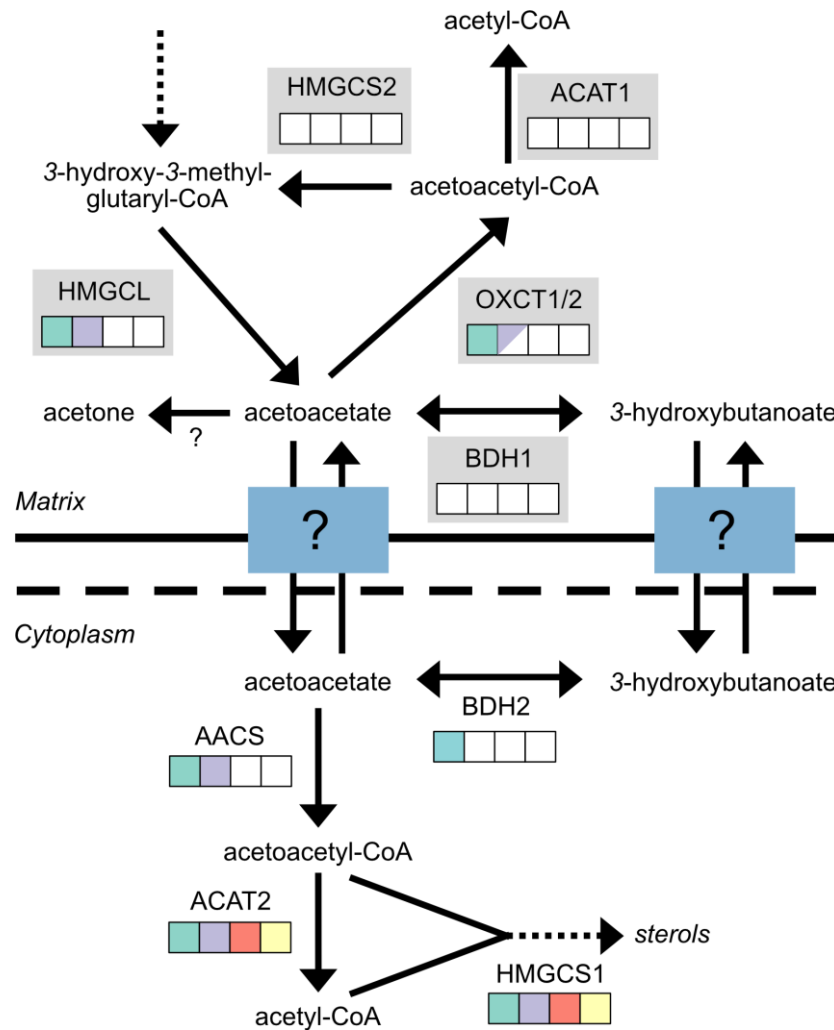


Figure 4.19 Phylogenetic profile of ketone body metabolism in platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Grey boxes mark IMPI positive training genes.

As study of the characterised carriers had shown that single genes could also be linked to the function of a transporter, I also looked at the four additional metabolic genes from non-enriched pathways which fit the phylogenetic profile: *GSTZ1*, *NUDT19*, *AASS*, and *SHMT2*.

- *GSTZ1* (glutathione *S*-transferase zeta 1) encodes a maleylacetoacetate isomerase, important in tyrosine/phenylalanine degradation. Though this was not identified as part of the enriched pathways, tyrosine degradation does lead to the production of acetoacetate – one of the three ketone bodies – linking this gene into other missing parts of metabolism.

- *NUDT19* is a renal CoA-diphosphohydrolase, involved in the control of coenzyme A (CoA) concentrations through its breakdown (Shumar *et al.* 2018). While there are many CoA-based compounds in the enriched pathways there is not an obvious link to potential transport substrates.
- *AASS* (aminoadipate-semialdehyde synthase) encodes a protein that catalyses the first two stages of the lysine degradation pathway, which I have previously explored (Figure 4.9). There is no obvious uncharacterised transport step in this pathway.
- *SHMT2* (serine hydroxymethyltransferase 2) encodes a protein which is part of mitochondrial glycine and tetrahydrofolate metabolism. It catalyses the conversion of glycine to serine via donation of a methyl group from methyltetrahydrofolate. The proteins forming the other mitochondrial parts of this pathway are not identified in any platyhelminthes, making it hard to speculate on a link to transport activity.

To further prioritise potential transport substrates, I investigated the literature on the mitochondrial carriers themselves, for evidence to support any of the potential transport substrates. Robinson *et al.* (2008) used the three-repeat structure of the mitochondrial carriers to investigate substrates of the transporters, by looking for symmetric and asymmetric residues between the three repeats. As most transported substrates are asymmetric, residues involved in the binding of substrates are also likely to be asymmetric between the three repeats; whereas symmetric residues are more likely to be involved in the transport mechanism. From this analysis, they predict that the substrate for metazoan *UCP5* (*SLC25A14*) and related transporters is relatively small, as there are few asymmetric residues around the predicted binding site. Arginine residues are located at positions in the binding site similar to the oxoglutarate and dicarboxylate transporters. Together, these features predict that *SLC25A14/30* may transport small keto acids.

Looking at the structures of the potential transport substrates predicted by the phylogenetic profiling (Figure 4.20), this would particularly support the idea of *SLC25A14/30* as possible transporters of the ketone body acetoacetate or the branched chain α -keto acids produced by the first step in the breakdown of the branched chain amino acids.

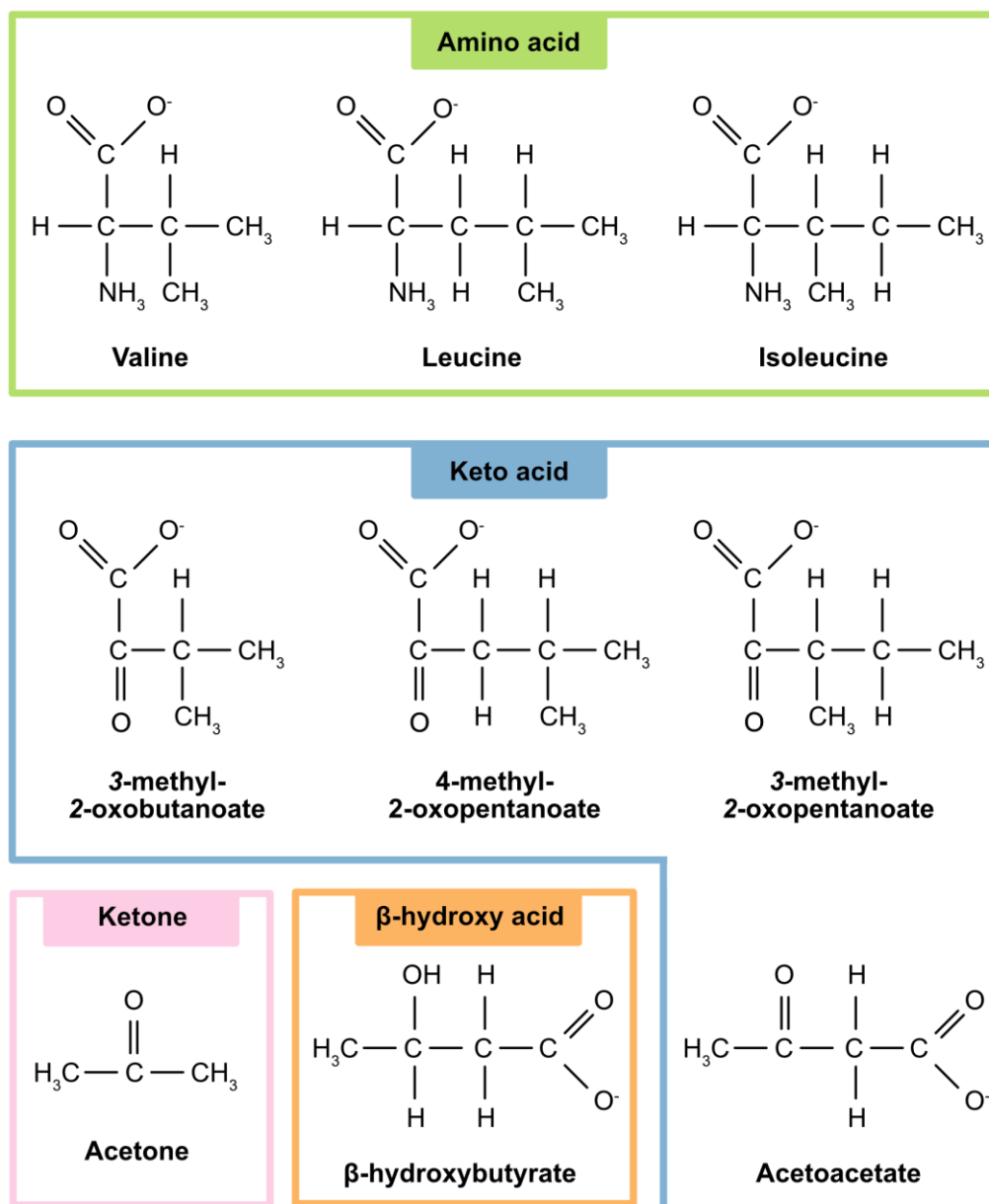


Figure 4.20 Structures of potential transport substrates for SLC25A14/30, grouped by the type of substrate.

mRNA expression analysis shows high levels of expression of *SLC25A14* in the brain and reproductive tissues (Yu *et al.* 2000; Thul *et al.* 2017), whereas *SLC25A30* mRNA is highly expressed in kidney and muscle tissues (Haguenauer *et al.* 2005; Thul *et al.* 2017). These are all tissues which are known to oxidise ketone bodies for use, after their production by the liver (Laffel 1999). This is particularly true of the brain which does not utilise fatty acids as an energy source when glucose concentrations are low, such as during fasting (Owen *et al.* 1967).

Other evidence in model organisms also suggests a link between this group of mitochondrial transporters and ketone bodies. Fruit flies (*Drosophila melanogaster*) with *SLC25A14* knocked out are more sensitive to the effects of fasting (Sánchez-Blanco *et al.* 2006). *SLC25A14* expression is also increased in rats fed on a synthetic ketone-ester diet (Kashiwaya *et al.* 2010) and in mice fed on a ketogenic diet (Ramsden *et al.* 2012).

In summary, *Table 4.9* shows a prioritised list of candidate transport substrates for the mitochondrial transporters *SLC25A14/30*, from the evidence produced by phylogenetic analysis and additional evidence from the literature.

Table 4.9 Prioritised candidate transport substrates from phylogenetic profiling for *SLC25A14/30* with types of associated evidence.

Rank	Compound(s)	Phylogenetic Pattern	Symmetry Analysis	Literature
1	Acetoacetate	X	X	X
2	Branched chain α -keto acid	X	X	
	Acetone, β -hydroxybutyrate	X		X
3	Branched chain amino acids	X		

Uncharacterised transporters: *SLC25A43*

SLC25A43 is an uncharacterised member of the mitochondrial carrier family which is not closely related to any other family members, at least in humans (*Figure 4.3*). In the studied platyhelminthes, predicted orthologues of *SLC25A43* are only identified in the free-living *S. mediterranea* (*Figure 4.4*), whereas in the studied nematodes, predicted orthologues are only identified in the *Dorylaimia* (*Figure 4.5*).

Fifteen IMPI 2017 positive training genes match or almost match the phylogenetic pattern of *SLC25A43* in the studied nematodes (*Appendix II – Table 5*). However, Reactome analysis does not identify any significant metabolic pathways and only one of the genes (*CYP1A1*) is clearly linked to metabolism. *CYP1A1* is a member of the cytochrome P450 family, involved in the metabolism of xenobiotic compounds and endogenous steroid metabolism (Lee *et al.* 2003). However, its role within the mitochondria is poorly defined.

Therefore, I moved on to investigating the genes fitting the phylogenetic pattern in the platyhelminthes. This is the same pattern as seen for the oxoadipate transporter *SLC25A21*, where lysine degradation was linked to transport function (*Figure 4.9*). I could not identify any additional uncharacterised transport steps in the lysine degradation pathway and associated metabolism, so I then considered the other metabolic genes with matching phylogenetic patterns.

Apart from the genes linked to lysine degradation, there are an additional forty-three genes with the same phylogenetic pattern as *SLC25A43* in the platyhelminthes, of which nineteen encode proteins involved in metabolism. Reactome pathway enrichment analysis identified ‘choline catabolism’ as an enriched pathway within this set of genes (*Table 4.2*). Choline catabolism occurs partially in the mitochondria and partially in the cytoplasm and is a source of methyl units (important in one carbon metabolism) through the methionine cycle. I investigated the phylogenetic patterns of choline catabolism and the cytoplasmic metabolism it is linked to (*Figure 4.21*).

The genes encoding proteins used to process methionine through to cysteine, allowing donation of a methyl-group from *S*-adenosyl methionine (SAM), are maintained across the studied platyhelminthes species. However, two routes which allow recycling of methionine from homocysteine (one dependent on choline, the other dependent on tetrahydrofolate metabolism) are lost. Choline dehydrogenase is the only gene of choline catabolism identified in a platyhelminth species beyond the free-living *S. mediterranea*. According to the known sub-cellular localisation of the enzymes involved in choline metabolism in mammals, dimethylglycine and either betaine or betaine aldehyde (all parts of choline metabolism) must cross the inner mitochondrial membrane, presumably via a carrier (*Figure 4.21*). Evidence suggests that this transporter is different from the choline transporter (*SLC44A1*), as the products of choline metabolism do not inhibit choline transport to a significant level (Porter *et al.* 1992). Therefore, dimethylglycine, betaine and betaine aldehyde are candidate transport substrates for *SLC25A43*.

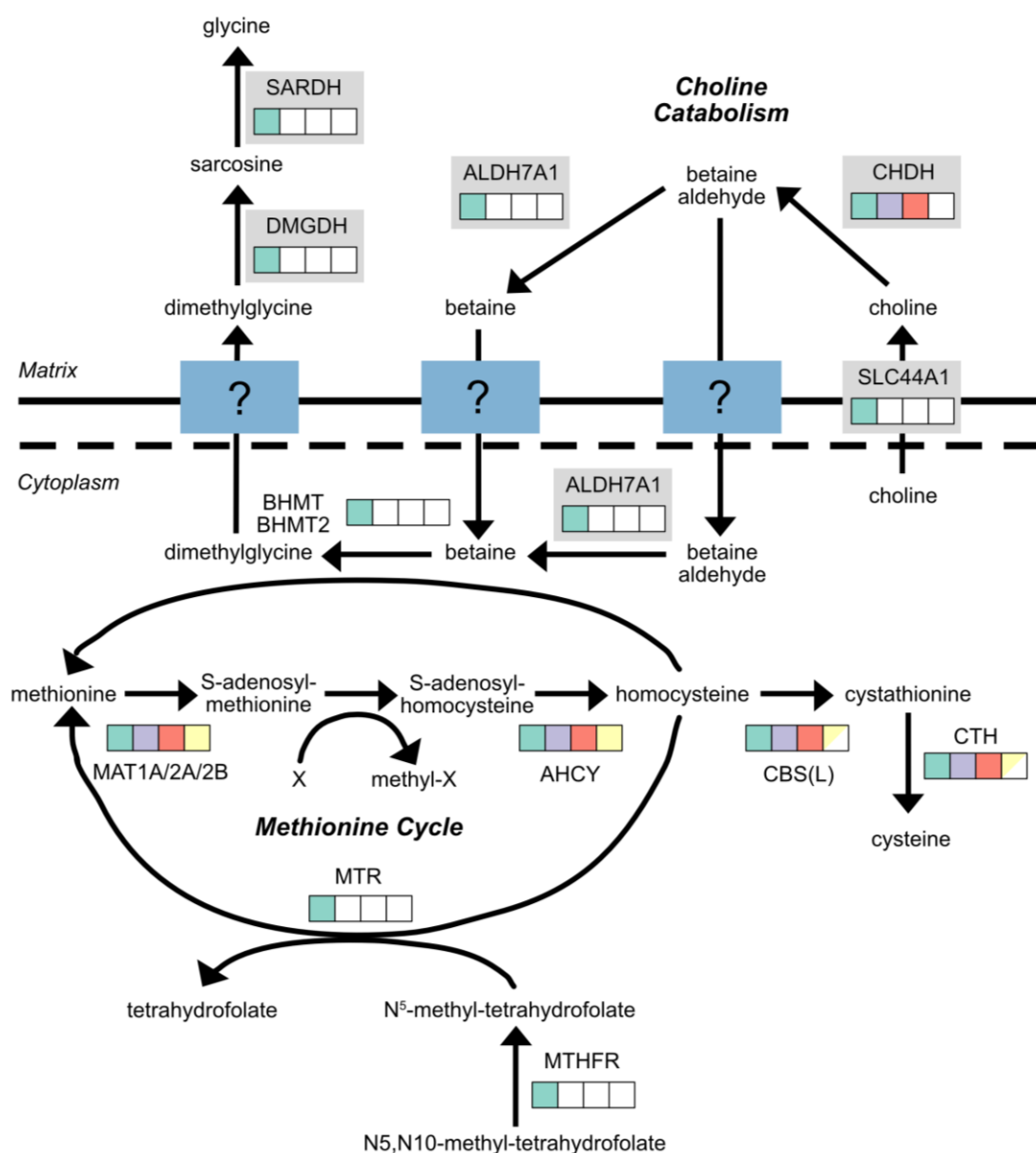


Figure 4.21 Phylogenetic profile of choline catabolism and the methionine cycle in the studied platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Grey boxes mark IMPI positive training genes. Half-filled boxes indicate this protein is missing in some of the species of that group.

Looking at the other metabolic genes with the same phylogenetic pattern as *SLC25A43*, another pathway of possible interest is GABA (γ -aminobutyric acid or 4-aminobutanoate) catabolism. GABA is a neurotransmitter which is produced in the cytoplasm or Golgi apparatus of the brain (and other tissues) from glutamate (Kanaani *et al.* 2015). GABA metabolism is linked to choline metabolism via betaine, as the plasma membrane transporter *BGT1* can transport both betaine and GABA into the cell (Yamauchi *et al.* 1992). GABA is

broken down to succinate semialdehyde in the mitochondrial matrix, which can then be converted to succinate and enter the TCA cycle. Genes for the two key reactions of GABA metabolism – cytoplasmic *GAD1/2* (glutamate decarboxylase) for synthesis and mitochondrial *ABAT* (4-aminobutanoate aminotransferase) for catabolism – are only identified in the free-living *S. mediterranea* (Figure 4.22), implying GABA metabolism is unique to this species in the studied platyhelminthes. Transport of GABA into the mitochondria is necessary for eventual catabolism, suggesting GABA as a candidate compound for transport by *SLC25A43*. An additional neurotransmitter, γ -hydroxybutyrate, is also metabolised to succinate semialdehyde in the mitochondria, by a gene only present in *S. mediterranea* (Figure 4.22). The endogenous production routes of γ -hydroxybutyrate are unclear, but this compound may also be a candidate for *SLC25A43* transport.

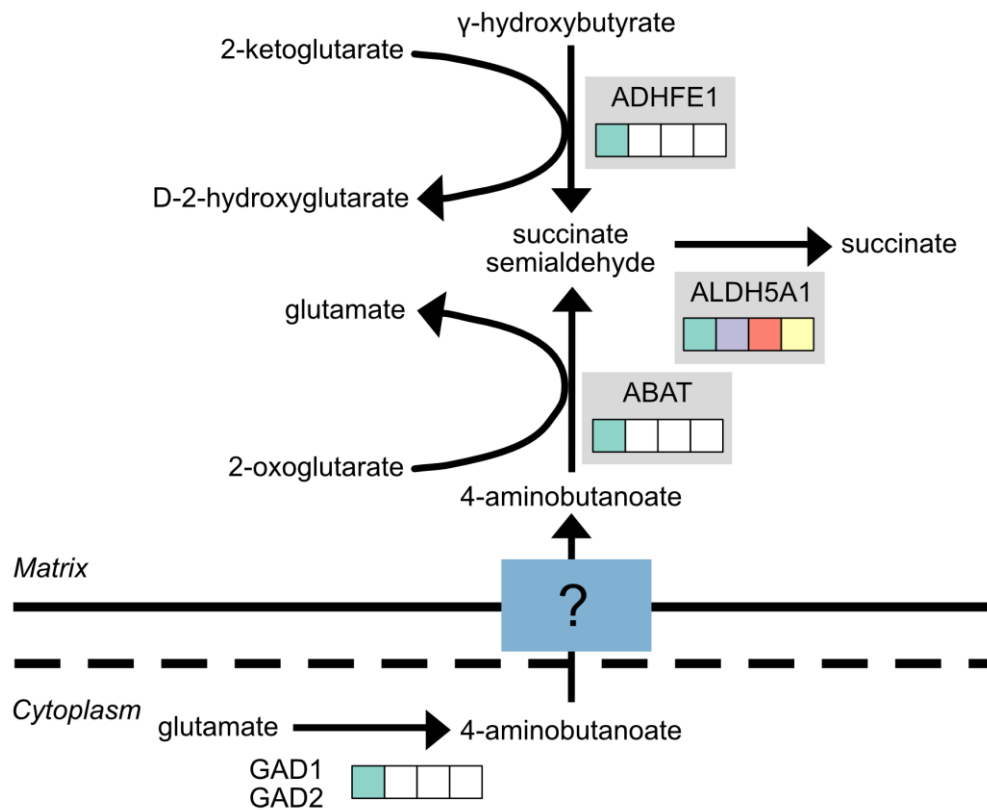


Figure 4.22 Phylogenetic profile of mitochondrial GABA (4-aminobutanoate) catabolism in studied platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Grey boxes mark IMPI positive training genes.

Another pathway of possible interest is the carnitine synthesis pathway (Figure 4.23).

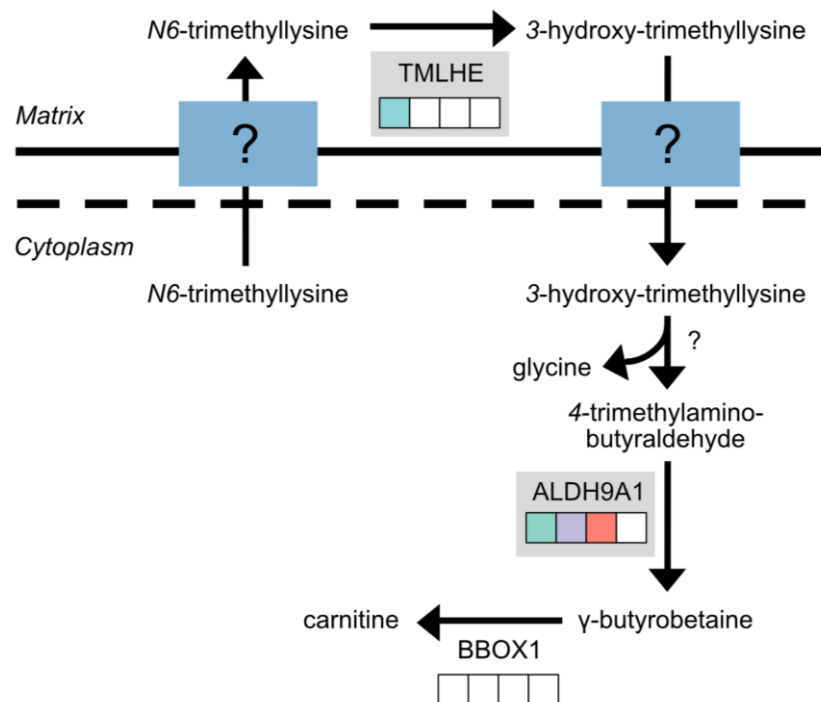


Figure 4.23 Phylogenetic profile of carnitine synthesis in studied platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Grey boxes mark IMPI positive training genes.

A single gene in this pathway is thought to encode a mitochondrial enzyme – *TMLHE* (trimethyllysine hydroxylase, epsilon). This enzyme catalyses the first step in the pathway, producing 3-hydroxy-trimethyllysine from *N6*-trimethyllysine. A predicted orthologue of this gene was only identified in *S. mediterranea*. *N6*-trimethyllysine is predicted to be sourced from the breakdown of proteins in the lysosome or proteasome (Strijbis *et al.* 2010) and so would need to enter the mitochondria.

The remaining steps of the pathway are thought to be cytoplasmic in humans, although the enzyme catalysing the second step has not been identified. Of the two remaining enzymes, *ALDH9A1* is identified in all platyhelminthes except liver flukes and *BBOX1* is not identified in any platyhelminthes. *ALDH9A1* is a promiscuous enzyme, so may be catalysing other reactions in the liver flukes and blood flukes. *BBOX1* may be truly missing from *S. mediterranea*, but as this analysis only looks at a single free-living species, it may also be missing due to poor protein annotation. Knowledge of the enzyme catalysing the second step of the pathway and its predicted platyhelminthes orthologues would help clarify this situation.

Without that information, both the substrate (*N*6-trimethyllysine) and product (3-hydroxy-trimethyllysine) of *TMLHE* are potential transport substrates for *SLC25A43*. γ -butyrobetaine is a potential substrate depending on subcellular localisation of the later steps and assuming carnitine synthesis occurs in *S. mediterranea*.

I also investigated each of the additional fourteen genes with links to metabolism showing the same phylogenetic pattern as *SLC25A43*, which were not in any enriched pathway (*Table 4.10*). None of these genes seemed to have links to a possible transported mitochondrial compound.

Table 4.10 Additional IMPI positive training set metabolic genes with phylogenetic profiles matching *SLC25A43* in the studied platyhelminthes.

Ensembl ID	Gene	Function
ENSG00000167315	<i>ACAA2</i>	Fatty acid β -oxidation
ENSG00000167107	<i>ACSF2</i>	Fatty acid synthesis
ENSG00000154930	<i>ACSS1</i>	Acetyl-CoA production from acetate
ENSG00000113492	<i>AGXT2</i>	Multifunctional aminotransferase
ENSG00000145439	<i>CBR4</i>	Fatty acid synthesis
ENSG00000132199	<i>ENOSF1</i>	Fuconate dehydratase
ENSG00000189221	<i>MAOA</i>	Outer membrane monoamine oxidase
ENSG00000103150	<i>MLYCD</i>	Control of fatty acid metabolism through controlling levels of malonyl-CoA
ENSG00000161653	<i>NAGS</i>	N-acetylglutamate production
ENSG00000036473	<i>OTC</i>	Urea cycle
ENSG00000128050	<i>PAICS</i>	Purine biosynthesis. Mainly cytoplasmic but some located in mitochondria
ENSG00000179598	<i>PLD6</i>	Possibly cardiolipin phosphatase, though located in the wrong membrane
ENSG00000148334	<i>PTGES2</i>	Prostaglandin E2 synthesis
ENSG00000128311	<i>TST</i>	Sulphur metabolism

The salt bridges at the cytoplasmic side of the transporter are predicted to be very weak, whereas the matrix side salt bridges look normal, suggesting *SLC25A43* may be an importer (Robinson *et al.* 2008). Looking at the symmetry analysis, the residues present in the binding pocket predict a possible nucleotide binding pocket, with positive residues to support binding the phosphate group. This is supported by the phylogenetic tree of human transporters (*Figure 4.3*), where *SLC25A43* groups with the nucleotide transporters and the CoA transporters (the structure of CoA includes a phosphoadenosine ring). None of the candidate transport substrates have structures which fit these criteria (*Figure 4.24*).

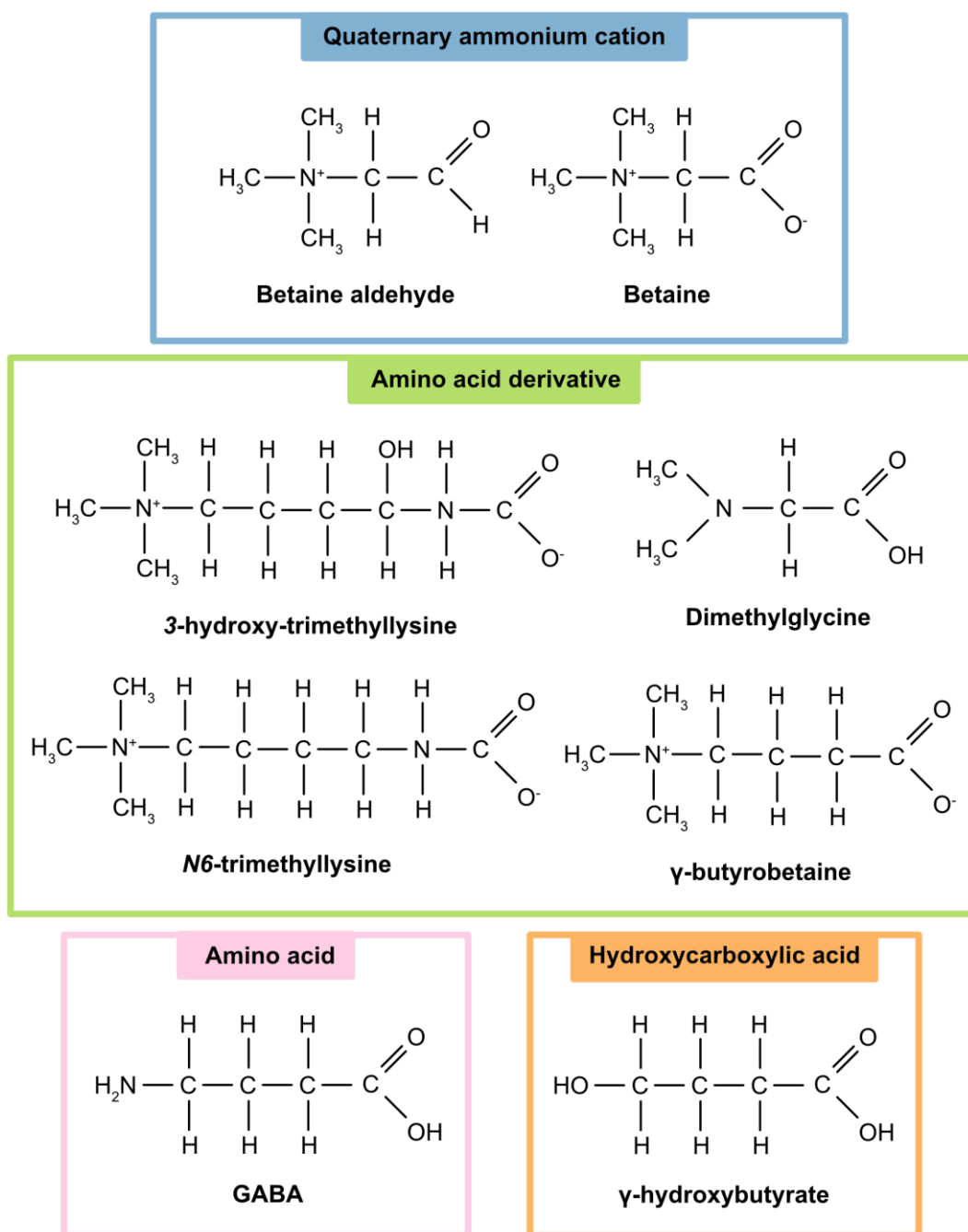


Figure 4.24 Structures of potential transport substrates for SLC25A43, grouped by the type of substrate.

There is very little literature on the carrier *SLC25A43*. Protein expression analysis by the Human Protein Atlas indicates the expression is highest in the lungs, liver, gastrointestinal tract and bone marrow (Uhlén *et al.* 2015); whereas mRNA expression analysis by another group shows high expression in the brain, adrenal gland and skeletal muscle (Haitina *et al.* 2006). Human Protein Atlas cellular localisation analysis predicts that *SLC25A43* is localised to the nuclear membrane, though this is only supported by one antibody (Thul *et al.* 2017).

This may be linked to a putative role for *SLC25A43* in cell cycle regulation, as seen in a breast cancer cell line (Gabrielson *et al.* 2016).

In summary, phylogenetic profiling of mitochondrial proteins in the nematodes and platyhelminthes does not identify a strong candidate for the transport substrate of *SLC25A43*. Evidence from the literature suggests that this transporter may be localised to the nuclear membrane, rather than the mitochondrial inner membrane, possibly transporting a nucleotide or related compound.

Uncharacterised transporters: SLC25A44

SLC25A44 is an uncharacterised member of the mitochondrial carrier family (*Figure 4.3*). While this carrier is present in all studied nematode species (*Figure 4.5*), it is missing specifically from the tapeworms in the studied platyhelminthes (*Figure 4.4*). Twenty-eight additional IMPI positive training set genes exhibit the same phylogenetic pattern, of which sixteen are clearly associated with metabolic processes (*Appendix II – Table 6*). Reactome pathway enrichment analysis of these genes identifies several overrepresented metabolic pathways in this gene set, of which the top three are listed in *Table 4.11*.

Table 4.11 Top three enriched Reactome metabolic pathways for IMPI genes only missing in the tapeworms, out of the studied platyhelminth species.

Reactome pathway	Count	<i>p</i> -value	Benjamini corrected <i>p</i> -value
Glycine degradation	2	5.68×10^{-5}	1.99×10^{-3}
Molybdenum cofactor synthesis	2	2.26×10^{-4}	3.83×10^{-3}
Pyruvate metabolism and Citric Acid (TCA) cycle	3	3.25×10^{-4}	4.55×10^{-3}

The glycine degradation pathway refers to the four genes encoding the glycine cleavage system: *AMT*, *DLD*, *GCSH* and *GLDC*. These genes encode proteins which catalyse the decarboxylation of glycine linked to the methylation of tetrahydrofolate – an important one carbon donor. Predicted orthologues of *AMT* and *GLDC* are not identified in the tapeworms, whereas predicted orthologues of *DLD* and *GCSH* are. *DLD* encodes a protein which is a

component of other dehydrogenase complexes, such as the pyruvate dehydrogenase complex, explaining its identification in tapeworms. However, there is no current additional characterised function for *GCSH*, though it has been speculated that *GCSH* may play additional roles in cell survival (Kikuchi *et al.* 2008). The presence of *GCSH* in the tapeworms supports this hypothesis. However, the phylogenetic loss of the glycine cleavage system in tapeworms does not provide any candidate transport substrates for *SLC25A44*. The glycine cleavage system does result in the production of 5,10-methylene tetrahydrofolate, but *SLC25A32* is already characterised as the mitochondrial tetrahydrofolate transporter (Titus & Moran 2000; McCarthy *et al.* 2004).

Another metabolic pathway identified from the enrichment analysis is ‘Pyruvate metabolism and Citric Acid (TCA) cycle’ (Table 4.11) and includes the genes *IDH3B*, *L2HGDH* and *FH*. As previously discussed, *FH* (fumarate hydratase) is present in tapeworms, though it is a different class of fumarate hydratase to that seen in other metazoans. *IDH3B* is one of the three subunits of isocitrate dehydrogenase, which usually functions as a heterotetramer, consisting of two *IDH3A* subunits, one *IDH3B* subunit and one *IDH3G* subunit. Both *IDH3A* and *IDH3G* are identified in tapeworms, and these two subunits can form heterodimers with some basal isocitrate dehydrogenase activity (Ehrlich & Colman 1983). This may allow a complete TCA cycle in tapeworms to function without *IDH3B*. *L2HGDH* is a metabolic repair enzyme, catalysing the production of α -ketoglutarate from the neurotoxin L-2-hydroxyglutarate, which is itself produced by promiscuous activity of malate dehydrogenase (Van Schaftingen *et al.* 2009). These genes do not form a coherent pathway and do not suggest any candidate transport substrates.

The remaining top pathway from the Reactome pathway enrichment analysis is ‘molybdenum cofactor synthesis’. Molybdenum cofactor (MoCo) coordinates molybdenum for use in enzymes that catalyse redox reaction (Mendel 2013). MoCo is synthesised *de novo* in humans, and the first stage of this synthesis is localised to the mitochondria, while all subsequent stages are cytoplasmic. *MOCS1* catalyses the synthesis of cyclic pyranopterin monophosphate (cPMP) from GTP in mitochondria. This is then exported to the cytoplasm to complete the remaining stages of MoCo synthesis. *MOCS1* and the genes encoding subsequent MoCo synthesis enzymes are lost in the studied tapeworms (Figure 4.25).

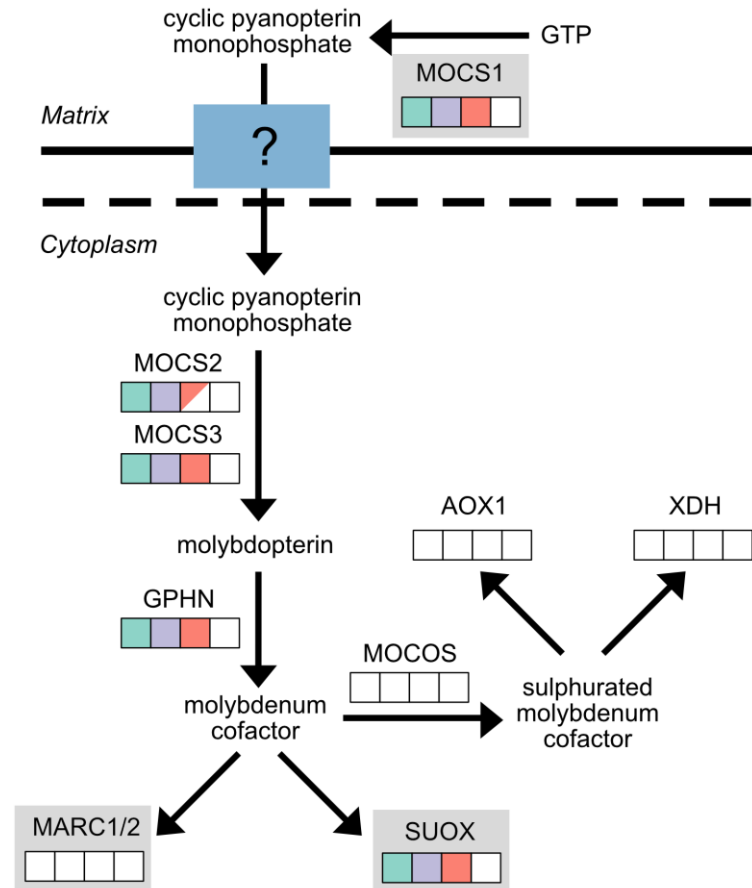


Figure 4.25 Phylogenetic profile of molybdenum cofactor biosynthesis in platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Half-filled boxes indicate this protein is missing in some of the species of that group. Grey boxes mark IMPI positive training genes.

This pattern is also reflected in the known MoCo dependent enzymes. Of the five identified in humans, only *SUOX* (sulphite oxidase) is identified at all in any of the studied platyhelminthes, and this is lost in the tapeworms (Figure 4.25).

A transporter (*ATM3*) identified in the plant species *Arabidopsis thaliana* has been linked to Fe-S and MoCo metabolism through changes noted in these pathways mutant plant studies (Teschner *et al.* 2010). Some have suggested that this gene may encode a cPMP transporter (Mendel 2013). However, further work in *A. thaliana* suggests that this transporter is not acting to move cPMP across the mitochondrial membrane, but may control the redox environment necessary for cPMP synthesis through the transport of glutathione (Kruse *et al.* 2018). The orthologous gene in humans is the mitochondrial *ABCB7*. Orthologues of this

gene are predicted in all the studied platyhelminthes, including tapeworms, supporting a role beyond molybdenum metabolism.

It has also been speculated that cPMP is hydrophobic enough to cross the mitochondrial inner membrane without a transporter (Mendel 2013). However, there has been no experimental evidence for this. The phylogenetic pattern analysis suggests that cPMP is a candidate substrate for the transporter *SLC25A44*.

After investigating the overrepresented pathways, I moved onto looking at the metabolism surrounding the remaining metabolic enzymes with the same phylogenetic pattern in platyhelminthes as *SLC25A44*. Another short synthesis pathway that transverses the mitochondrial membrane is creatine synthesis, which starts from glycine in the mitochondria (Figure 4.26). *GATM* (glycine amidinotransferase) which catalyses the formation of guanidinoacetate from glycine in the matrix fits the phylogenetic profile. However, the cytoplasmic enzyme *GAMT* (guanidinoacetate N-methyltransferase), which transforms guanidinoacetate to creatine, after movement across the inner mitochondrial membrane, does not. Phylogenetic profiling, therefore, does not support guanidinoacetate a candidate substrate for *SLC25A44*.

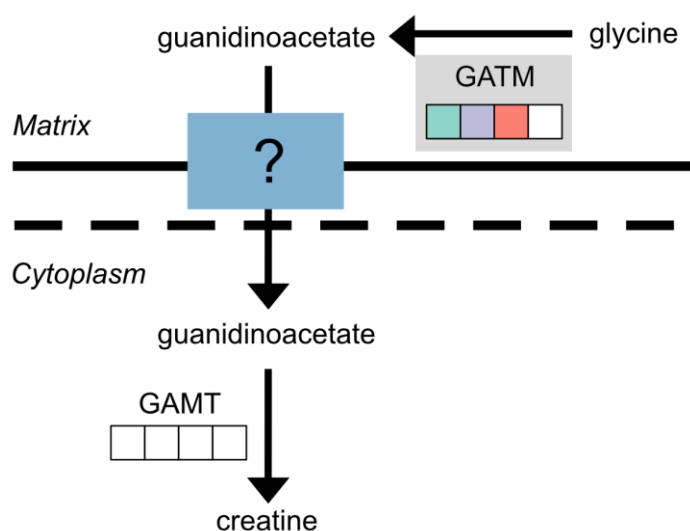


Figure 4.26 Phylogenetic profile of creatine biosynthesis in platyhelminthes. Coloured boxes indicate that the gene is present in the following platyhelminth species: green = free-living; purple = liver flukes; red = blood flukes; yellow = tapeworms. Grey boxes mark IMPI positive training genes.

The remaining metabolic genes identified from phylogenetic profiling (Table 4.12) did not appear to be linked to any potential *SLC25A44* transport compounds.

Table 4.12 Additional IMPI metabolic genes with phylogenetic profiles matching SLC25A44 in the studied platyhelminthes.

Ensembl ID	Gene	Function
ENSG00000059573	<i>ALDH18A1</i>	Proline, ornithine and arginine synthesis
ENSG00000143149	<i>ALDH9A1</i>	GABA metabolism
ENSG00000016391	<i>CHDH</i>	Choline metabolism
ENSG00000167113	<i>COQ4</i>	Coenzyme Q synthesis
ENSG00000102967	<i>DHODH</i>	Pyrimidine biosynthesis (outer side of inner mitochondrial membrane)
ENSG00000180185	<i>FAHD1</i>	Oxaloacetate to pyruvate (reverse of <i>PC</i>)
ENSG00000174099	<i>MSRB3</i>	Sulphur metabolism
ENSG00000173599	<i>PC</i>	Pyruvate to oxaloacetate (reverse of <i>FAHD1</i>)

Symmetry analysis suggests *SLC25A44* carries a substrate very different from other transporters (Robinson *et al.* 2008). Residues at substrate contact points include:

- an arginine, supporting a substrate with a charged group, such as a carboxylate or phosphate group (interacting with the positively charged arginine);
- and a tyrosine, supporting a substrate with a hydrophobic or uncharged polar side chain (interacting with the tyrosine ring).

The structure of cPMP (Figure 4.27) fits this description, as it contains both a charged phosphate group, and a complex pyranopterin ring structure. However, symmetry analysis also suggests that the carrier will function as both an importer and exchanger, whereas cPMP should be exported. cPMP may be exported in exchange for another substrate.

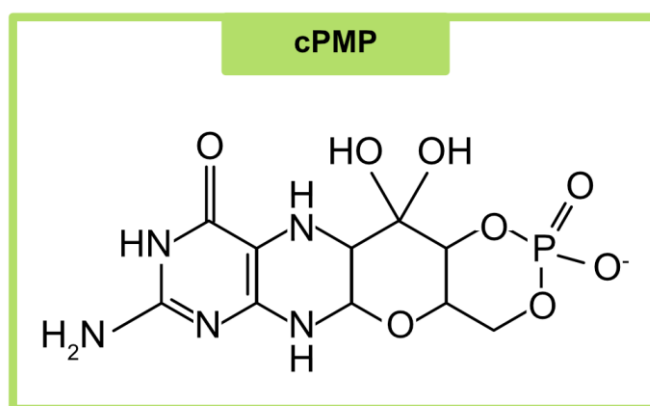


Figure 4.27 Structure of cyclic pyranopterin monophosphate (cPMP) – a potential transport substrate for SLC25A44.

As with *SLC25A43*, there is very little literature investigating the *SLC25A44* carrier. mRNA expression analysis suggests that *SLC25A44* is highly expressed in the brain and kidney, though there is expression across a range of tissues (Haitina *et al.* 2006; Uhlén *et al.* 2015). The MoCo synthesis genes are also widely expressed across tissues.

In summary, phylogenetic profiling in the platyhelminthes and symmetry analysis supports cPMP as a possible candidate transport substrate for the uncharacterised mitochondrial carrier *SLC25A44*.

Uncharacterised transporters: SLC25A45/47/48

SLC25A45, *SLC25A47* and *SLC25A48* are all uncharacterised mitochondrial carriers, which cluster together in the human phylogenetic tree (*Figure 4.3*). In the studied platyhelminthes, orthologues of this group of carriers are not predicted in the tapeworms or the blood fluke *Schistosoma japonicum*. This type of phylogenetic pattern, with loss in only one of several closely related species of schistosomes, provides two possibilities: either this is realistic, and this carrier group has been specifically lost in *S. japonicum* but not in the other studied blood flukes, or the lack of identification of an orthologue in *S. japonicum* is due to incomplete sequencing or annotation at the genome level. In the latter case, predicted transport substrates would be the same as for *SLC25A44*.

I also considered the former case, by identifying IMPI positive training genes with the same exact phylogenetic pattern as *SLC25A45/47/48*. There are four additional genes with a matching phylogenetic pattern, of which two have characterised roles in metabolism (*Table 4.13*). Examining the metabolism around these two enzymes did not provide any useful information regarding possible carrier function.

Table 4.13 Additional IMPI genes with phylogenetic profiles matching *SLC25A45/47/48* in the studied platyhelminthes.

Ensembl ID	Gene	Function
ENSG00000196072	<i>BLOC1S2</i>	Apoptosis
ENSG00000134463	<i>ECHDC3</i>	Unknown
ENSG00000115419	<i>GLS</i>	Glutamine catabolism
ENSG00000124370	<i>MCEE</i>	Branched chain amino acid catabolism

In the studied nematodes, predicted orthologues of *SLC25A46/47/48* are missing in the species forming the Filarioidea and the Dorylaimia (*Figure 4.5*). Thirty-two IMPI positive training genes match this phylogenetic pattern, with an allowance for lack of identification of an orthologue in one additional species. Of these, twenty can be linked to some type of metabolic process. Only one specific metabolic pathway is overrepresented in these genes on Reactome pathway analysis – ‘lysine catabolism’ (Benjamini corrected p -value = 4.73×10^{-3}). I had previously investigated lysine catabolism in the nematodes when exploring the oxoadipate carrier (*Figure 4.9*) and did not identify any additional necessary transport steps.

I also investigated the other seventeen metabolic genes which were not included in an enriched pathway (*Table 4.14*) but did not identify any good candidates for *SLC25A45/47/48* transported compounds.

Table 4.14 Additional IMPI metabolic genes with phylogenetic profiles matching *SLC25A45/47/48* in the studied nematodes.

Ensembl ID	Gene	Function
ENSG00000197150	<i>ABCB8</i>	Iron transport
ENSG00000111271	<i>ACAD10</i>	Fatty acid β -oxidation
ENSG00000151498	<i>ACAD8</i>	Valine catabolism
ENSG00000004455	<i>AK2</i>	Adenylate kinase isoform
ENSG00000136010	<i>ALDH1L2</i>	Tetrahydrofolate metabolism
ENSG00000164904	<i>ALDH7A1</i>	Choline metabolism
ENSG00000242110	<i>AMACR</i>	Fatty acid β -oxidation
ENSG00000019186	<i>CYP24A1</i>	Vitamin D3 metabolism
ENSG00000104325	<i>DECR1</i>	Fatty acid β -oxidation
ENSG00000105607	<i>GCDH</i>	Lys/Trp degradation
ENSG00000119927	<i>GPAM</i>	Glycerolipid synthesis
ENSG00000100577	<i>GSTZ1</i>	Phe/Trp degradation
ENSG00000106049	<i>HIBADH</i>	Branched chain amino acid degradation
ENSG00000189221	<i>MAOA</i>	Monoamine oxidation
ENSG00000213965	<i>NUDT19</i>	CoA diphosphohydrolase
ENSG00000110435	<i>PDHX</i>	Pyruvate dehydrogenase complex
ENSG00000185973	<i>TMLHE</i>	Carnitine biosynthesis

Therefore, phylogenetic profiling in platyhelminthes and nematodes did not identify any good potential candidate transport substrates for transport by this group of mitochondrial carriers.

Discussion

Mitochondrial carrier transport substrate identification has previously been supported by genetic and biochemical methods (Gutiérrez-Aguilar & Baines 2013), as well as bioinformatic analysis looking at protein sequences (Robinson *et al.* 2008). However, there are still several members of the human mitochondrial carrier family which have not been functionally characterised. Phylogenetic profiling provides another analytical technique to help support mitochondrial carrier functional annotation, by linking metabolic pathways to carriers through their presence and absence in a set of related species (Loganantharaj & Atwi 2007). This is dependent on finding a good group to analyse, with enough sequenced divergent species and informative phylogenetic information related to the transporters of interest (Jothi *et al.* 2007) and producing a well-defined set of gene orthologues (Snitkin *et al.* 2006).

Therefore, I identified two Metazoan phyla (Nematoda and Platyhelmintha) which include a variety of parasitic species, with variable loss of metabolic pathways within each taxon. Metazoans were chosen in an attempt to reduce the amount of alternative metabolic pathways present compared to human metabolism, which may otherwise confuse a phylogenetic analysis which is based primarily on human genes and pathways (Jothi *et al.* 2007). Where possible, I included several related species to counter annotation-based limitations, as it has previously been shown that there is reduced predictive performance in large-scale phylogenetic profiling in less well-annotated species (Jothi *et al.* 2007). The chosen phyla maintain a large proportion of the mitochondrially localised metabolism seen in humans (approximately 67% in each phylum) and show variation in the identification of the mitochondrial carriers, including some characterised and some uncharacterised members (Figures 4.4 and 4.5).

Analysis of the characterised carriers which showed interesting phylogenetic profiles provided good test cases for phylogenetic profiling within the chosen phyla. Where there was an interesting phylogenetic pattern for a carrier in both the platyhelminthes and the nematodes, there were not consistent results in the analysis of the surrounding metabolism between the phyla. This is consistent with the idea of a greater numbers of alternative metabolic pathways within the eukaryotes (Jothi *et al.* 2007), which may differ between the nematodes and platyhelminthes. For *SLC25A21* (the oxoadipate carrier), there was a strong link between the phylogenetic profile of the carrier and the genes encoding the surrounding lysine/tryptophan catabolism in the platyhelminthes (Figure 4.9), but not in the nematodes.

This was the only characterised carrier with such a strong link and suggests that phylogenetic profiling may be useful where there is well-characterised metabolism around the transport substrate. It also indicates that, at least in the Metazoa, it may be best to consider taxa separately when using phylogenetic profiling for transport substrate identification.

Other characterised transporters that were investigated did not provide such well-defined results. In some cases, there were potential links between the patterns of single genes with the function of the transporter (e.g. *SLC25A38* in the platyhelminthes). These cases would be much harder to identify for uncharacterised transporters without pre-knowledge of what pathways to look at. However, this suggested that it was worth considering each individual metabolic gene which fits the phylogenetic pattern. In other cases, there was no obvious phylogenetic link between the transport substrate and surrounding metabolism, supporting the idea that a variety of different techniques will be necessary to characterise all the remaining mitochondrial transporters. These lessons were then taken forward to the investigation of the uncharacterised transporters.

Four groups of uncharacterised transporters were identified with interesting phylogenetic profiles in the platyhelminthes, the nematodes or both (*Figures 4.4 – 4.5*). Analysis of genes with close phylogenetic profiles produced sensible candidate transport substrates for two of these carrier groups, both using the phylogenetic patterns observed in the platyhelminthes. The lack of sensible candidate substrates for the remaining two uncharacterised carrier groups points to the limitations of this technique. For *SLC25A43*, the success of the analysis may be limited by the localisation of the transporter. If the protein encoded by this gene is localised to the nuclear membrane, as is suggested by subcellular localisation analysis in the Human Protein Atlas (Thul *et al.* 2017), mitochondrial metabolism would not be expected to link to any transport substrate.

For the mitochondrial carrier group consisting of two uncoupling protein related genes (*SLC25A14* and *SLC25A30*), phylogenetic profiling provided several candidate transport substrates (*Table 4.8*). Using supporting information from the literature, I prioritised these candidates, with the ketone body acetoacetate being the best-supported prediction. Acetoacetate fits the predicted substrate type of a small keto acid, produced from analysis of carrier sequence symmetry (Robinson *et al.* 2008). This prediction is also supported by the links between the carriers *SLC25A14/30* and ketone body metabolism in several different

parts of the scientific literature surrounding these carriers and in the tissue expression patterns (Sánchez-Blanco *et al.* 2006; Kashiwaya *et al.* 2010; Ramsden *et al.* 2012).

The second uncharacterised carrier with a good candidate substrate from phylogenetic profiling is *SLC25A44*, for which the molybdenum coenzyme precursor cPMP was identified as a possible transport substrate (*Figure 4.25*). *SLC25A44* does not group closely with any other member of human mitochondrial carrier family (*Figure 4.3*) and contains a unique tyrosine in the substrate binding region of the carrier (Robinson *et al.* 2008). While these characteristics make the transporter particularly interesting, they also make identifying possible transport substrates to fit these characteristics more difficult, as there are no similar transporters to compare substrates with. Predictions are also limited by the lack of literature around this carrier. cPMP as a potential substrate fits well with what has been suggested about the transporter. The structure of cPMP is not similar to any transport substrate of the characterised transporters, and the flat rings in cPMP (*Figure 4.27*) would be predicted to interact with the aromatic ring in the tyrosine structure.

These results show the potential utility of phylogenetic profiling within the eukaryotes to support the characterisation of transporters, but also the limitations. Profiling techniques are dependent on finding well-characterised species with variation in the genes of interest, and on the consistency of metabolic pathways between the species being compared. Sequencing of additional species from other metazoan groups containing parasitic or highly derived species (such as parasitic jellyfish and other parasitic species within the Cnidaria) may provide additional sources of interesting phylogenetic patterns for the characterisation of human proteins. Knowledge of the loss of pathways within certain types of species may also provide support for a targeted phylogenetic analysis, looking for the transporter of a specific substrate. For example, vitamin B12 metabolism appears to have been lost in true insects, but not in other members of the Pancrustacea. Therefore, these species may be a good source of data to further characterise this metabolic pathway. This technique could also be used to prioritise transport substrates for other groups of transporters beyond the mitochondrial carriers, particularly where their localisation in the cell is known (e.g. mitochondrial membrane, nuclear membrane).

Phylogenetic profiling will not prove informative in all cases. Even in those cases where the technique does produce sensible predictions, further work is still necessary to prove or disprove the predictions. Literature evidence and additional bioinformatic studies, such as

those looking at the structure of the transporters in this case, can help further prioritise candidates. Eventually, however, the predictions must be tested using either genetic or biochemical techniques, such as transport assays (Gutiérrez-Aguilar & Baines 2013) and this will be the true test of the quality of the predictions. Early work has been started for testing the ability of the *SLC25A14/30* carriers to transport ketone bodies.

Conclusions

In this chapter, I identified two Metazoan phyla (Nematoda and Platyhelmintha) to target for the purpose of predicting possible transport substrates of uncharacterised members in the mitochondrial carrier family. These phyla contain species with variation in the presence of metabolic pathways and mitochondrial carriers. I created curated datasets of orthologues of mitochondrial genes for sequenced members of these two phyla, to create a phylogenetic profile of presence and absence for each gene.

From test cases using characterised members of the mitochondrial carrier family, I observed a link between the phylogenetic pattern of a carrier and the surrounding metabolism in one case (*SLC25A21* in platyhelminthes), possible phylogenetic links between single genes and carriers in other cases, and no links in others. This supported the use of phylogenetic profiling to identify transporter function, but suggested results should not be expected for all uncharacterised transporters. Differences in metabolism were often not conserved between the two studied phyla, leading further analyses in the two phyla to be considered separately.

I then moved on to testing four groups of uncharacterised carriers. Of these four carrier groups, there were strong candidate transport substrates identified for two: *SLC25A14/30* as potential transporters of ketone bodies, which is also well-supported by the surrounding observational literature and *SLC25A44* as a potential transporter of the molybdenum coenzyme precursor cPMP. These cases provide justification for testing of transport activity by these carriers for the predicted substrates, using genetic or biochemical techniques.

Chapter 5

Exploring the history and function of
genes causing monogenetic mitochondrial
diseases

Introduction

Mitochondrial disease

Disruption of normal mitochondrial function is associated with a wide assortment of diseases. These range from common, multifactorial diseases often associated with age, such as Alzheimer's disease (Onyango *et al.* 2017), cancer (Ahn & Metallo 2015) and heart disease (Huss & Kelly 2005); to rare, monogenetic diseases often seen in paediatric patients, such as respiratory chain deficiencies (DiMauro & Schon 2003). In this chapter I use 'mitochondrial disease' to refer to the less common, monogenetic forms of disease, associated with mutations in genes encoding the mitochondrial proteome.

Mitochondrial diseases can be caused by mutations in mitochondrial DNA (mtDNA), but human mtDNA only encodes 37 genes, including 13 protein-encoding genes. Other disease-causing mutations occur in the nuclear DNA, which encode genes necessary for mitochondrial function, particularly the approximately 1,500 genes which encode the mitochondrial proteome (Meisinger *et al.* 2008). Although individually most mitochondrial diseases are rare, together they are predicted to affect up to 9 in every 100,000 children (Darin *et al.* 2001; Skladal *et al.* 2003) and 1 in 4,300 adults (Gorman *et al.* 2015). This number can be higher in certain populations, depending on the frequency of disease-associated alleles (Skladal *et al.* 2003).

Exome sequencing is an important tool in the genetic diagnosis of mitochondrial disease patients. However, the average success rate in larger scale studies is only around 50% (Calvo *et al.* 2012; Taylor *et al.* 2014). One bottleneck that may limit the success of exome sequencing is the prioritisation of candidate genes with rare variants for further study. Functional analysis of genes in cell lines or model organisms is time-consuming and resource-intensive. Therefore, where there is no obvious candidate, bioinformatic analysis is an important tool to identify and prioritise the most interesting genes to take forward into laboratory work (Moreau & Tranchevent 2012).

Features of disease genes

Understanding the features of known disease genes can inform prioritisation of candidate disease genes. Potentially informative features of disease genes have been identified by investigating phylogeny, orthology and evolution.

Mendelian disease genes:

- are under stronger purifying selection than those not associated with disease (Blekhman *et al.* 2008) and have a wider phylogenetic spread (i.e. orthologues are found in a wider range of taxa) (López-Bigas & Ouzounis 2004).
- are more likely to be of ancient origin (Domazet-Lošo & Tautz 2008; Maxwell *et al.* 2014), though genes associated with different types of diseases show variation in origin (Maxwell *et al.* 2014). For example, genes involved in diseases of nervous system development are more likely to originate post-split of the Metazoa from other eukaryotes (Maxwell *et al.* 2014).
- can be traced back to historical duplications, with genes without identified duplicates in humans more likely to be associated with disease (Dickerson & Robertson 2012). Genes with highly conserved human paralogues are less likely to be associated with disease (López-Bigas & Ouzounis 2004).

The evolutionary history of genes can be used to distinguish genes associated with different types of disease inheritance – recessive disease genes (where two faulty copies are necessary for a disease phenotype) and dominant disease genes (where only one faulty copy can lead to a disease phenotype). Dominant disease genes have, on average, a greater number of paralogues than recessive disease genes (Kondrashov & Koonin 2004; Furney *et al.* 2006); and are also more conserved between human and mouse, under stronger purifying selection (Furney *et al.* 2006; Blekhman *et al.* 2008; Cai *et al.* 2009). Cai *et al.* (2009) suggest that dominant disease genes are younger than recessive disease genes on average, though this was not statistically tested and the authors only looked as far back as the Metazoa, plus *Saccharomyces cerevisiae*.

I hypothesised that mitochondrial disease genes might show a distinct origin and phylogenetic pattern compared to other human disease genes, as the mitochondria have a unique

evolutionary history compared to the rest of the cell and are the site of a variety of essential and unique functions, such as the electron transport chain.

Chapter summary

In this chapter, I investigate the phylogenetic spread and origin of disease genes in IMPI compared to those not associated with disease. I use the same types of analyses to compare mitochondrial disease genes where the diseases are inherited recessively and dominantly; and look for any function associated with identified differences. Finally, I investigate the potential utility of studying naturally occurring human loss-of-function homozygotes and knockout lines of three model organisms (*Mus musculus*, *S. cerevisiae* and *Escherichia coli*), relative to prioritising candidate mitochondrial disease genes.

Methods

Defining a list of diseases of the mitochondrion

The list of human genes which form IMPI 2017 was used as the definition of the mitochondrial proteome, with use of the IMPI positive training set (a manually curated set of mitochondrial genes) as a more stringent dataset. This included 1,550 genes in IMPI 2017 and 1,130 genes in the IMPI positive training set.

The majority of gene-disease associations were retrieved from OMIM (Online database of Mendelian Inheritance in Man) (Amberger *et al.* 2015). The mode of disease inheritance was identified from OMIM, and assigned per gene as either recessive, dominant, both recessive and dominant, X-linked or mitochondrial. Additional monogenetic disease associations were retrieved from the literature, using the search pattern “‘{gene name}’ disease’ in Google Scholar. This method retrieved diseases associated with an additional 28 genes of the mitochondrial proteome (*Table 5.1*).

Definitions of taxa

Some analysis involves dividing the studied species into several taxa. These are defined below, and the assigned taxon of each studied species is listed in *Appendix I – Table 1*.

- Holozoa – animals and the closely related single-celled ichthyosporea and choanoflagellates
- Holomycota – fungi and *Fonticula*
- Amoebozoa – amoeboid protists
- Archaeplastida – red algae, green algae, land plants and glaucophytes
- SAR group – stramenopiles, alveolates and rhizaria
- Excavata plus – variety of unicellular species
- Archaea
- Bacteria

Table 5.1 IMPI 2017 genes associated with monogenetic diseases which were absent from OMIM and identified from a literature search.

Ensembl ID	Gene Name	Reference
ENSG00000138138	<i>ATAD1</i>	(Ahrens-Nicklas <i>et al.</i> 2017)
ENSG00000183978	<i>COA3</i>	(Ostergaard <i>et al.</i> 2015)
ENSG00000162377	<i>COA7</i>	(Lyons <i>et al.</i> 2016)
ENSG00000132591	<i>ERAL1</i>	(Chatzispyrou <i>et al.</i> 2017)
ENSG00000267673	<i>FDX1L</i>	(Spiegel <i>et al.</i> 2013)
ENSG00000164347	<i>GFM2</i>	(Fukumura <i>et al.</i> 2015)
ENSG00000115541	<i>HSPE1</i>	(Bie <i>et al.</i> 2016)
ENSG00000262814	<i>MRPL12</i>	(Serre <i>et al.</i> 2013)
ENSG00000125445	<i>MRPS7</i>	(Menezes <i>et al.</i> 2015)
ENSG00000133026	<i>MYH10</i>	(Tuzovic <i>et al.</i> 2013)
ENSG00000186010	<i>NDUFA13</i>	(Angebault <i>et al.</i> 2015)
ENSG00000189043	<i>NDUFA4</i>	(Pitceathly <i>et al.</i> 2013)
ENSG00000244005	<i>NFS1</i>	(Farhan <i>et al.</i> 2014)
ENSG00000178694	<i>NSUN3</i>	(Van Haute <i>et al.</i> 2016)
ENSG00000162396	<i>PARS2</i>	(Sofou <i>et al.</i> 2015)
ENSG00000232838	<i>PET117</i>	(Renkema <i>et al.</i> 2017)
ENSG00000147403	<i>RPL10</i>	(Brooks <i>et al.</i> 2014)
ENSG00000134419	<i>RPS15A</i>	(Ikeda <i>et al.</i> 2017)
ENSG00000181035	<i>SLC25A42</i>	(Shamseldin <i>et al.</i> 2016)
ENSG00000014824	<i>SLC30A9</i>	(Perez <i>et al.</i> 2017)
ENSG00000146350	<i>TBC1D32</i>	(Adly <i>et al.</i> 2014)
ENSG00000113272	<i>THG1L</i>	(Edvardson <i>et al.</i> 2016)
ENSG00000164983	<i>TMEM65</i>	(Nazli <i>et al.</i> 2017)
ENSG00000126602	<i>TRAP1</i>	(Saisawat <i>et al.</i> 2014)
ENSG00000170855	<i>TRIAP1</i>	(Poole <i>et al.</i> 2016)
ENSG00000043514	<i>TRIT1</i>	(Kernohan <i>et al.</i> 2017)
ENSG00000184470	<i>TXNRD2</i>	(Sibbing <i>et al.</i> 2011)
ENSG00000116874	<i>WARS2</i>	(Musante <i>et al.</i> 2017)

Phylostratigraphic analysis

Phylostratigraphic analysis was carried out as in Domazet-Lošo & Tautz 2008, except that some phylostrata were combined together in comparison to the original paper.

Euarchontoglires, Boreoeutheria and Eutheria were combined, as were Osteichthyes with Vertebrata, Olfactores with Deuterostomia, and Eumetazoa with Metazoa. Genes were assigned to a phylostratum based on the first of the defined phylostrata to contain a predicted orthologue of the gene, from the previously described orthologue dataset (*Chapter 2*).

Log odds ratio (OR) was calculated, using natural logarithms, as:

$$\log(OR) = \log\left(\frac{\text{no. of outcome 1 in phylostrata}/\text{no. of outcome 2 in phylostrata}}{\text{no. of outcome 1 not in phylostrata}/\text{no. of outcome 2 not in phylostrata}}\right)$$

where for comparing disease and non-disease, ‘outcome 1’ is disease and ‘outcome 2’ is no disease; and for comparing recessive and dominant inheritance, ‘outcome 1’ is dominant inheritance and ‘outcome 2’ is recessive inheritance.

Gene annotation and enrichment

GO Biological Process/KEGG pathway annotation and enrichment was carried out using the DAVID functional annotation tool (Huang *et al.* 2009a). Enriched clusters were identified using DAVID functional annotation clustering with medium classification stringency.

GO annotation (GO:0003824 ‘catalytic activity’ and GO:0008152 ‘metabolic process’) was assigned to genes using QuickGO (Binns *et al.* 2009). Genes of central mitochondrial metabolism were taken from the MitoCore model (Smith *et al.* 2017).

Essential genes in model organisms

Mouse (*Mus musculus*) genes annotated with the mammalian phenotype terms “complete prenatal lethality” (MP:0011091) and “complete perinatal lethality” (MP:0011089) were retrieved from MouseMine, as were mouse orthologues of human IMPI genes with no assigned mouse phenotype (Motenko *et al.* 2015). Human IMPI 2017 genes were mapped to corresponding mouse genes by name where possible and by manually assessing homology using data from the MitoMiner database (Smith & Robinson 2016) and BLASTp searches (Altschul *et al.* 1990, 1997) where not.

Yeast (*Saccharomyces cerevisiae*) essential genes were downloaded from the *Saccharomyces* Genome Deletion Project (Giaever *et al.* 2002)

(http://www-sequence.stanford.edu/group/yeast_deletion_project/Essential_ORFs.txt). The yeast strains used in this project were based on a derivative of the *S. cerevisiae* S288c strain, which was also used in the orthology dataset predictions, allowing mapping of gene names

between the two. Essential yeast genes were mapped to the corresponding human gene by comparing the predicted yeast orthologues of each human IMPI 2017 gene to the list of yeast essential genes.

Escherichia coli essential genes were downloaded from the EcoGene resource (Zhou *et al.* 2013) (http://www.ecogene.org/old/topic.php?topic_id=5). Essential *E. coli* genes were mapped to the corresponding human genes by comparing the identified *E. coli* orthologues of each IMPI 2017 gene to the list of *E. coli* essential genes.

Human loss-of-function homozygotes

Four loss-of-function (LoF) datasets were used to identify LoF genes in healthy humans:

- A dataset of sequenced and imputed LoF homozygotes and compound heterozygotes from a healthy Icelandic population (Sulem *et al.* 2015). This identified 105 IMPI genes with at least one imputed LoF homozygote or compound heterozygote and with minor allele frequency below 1%.
- A dataset of LoF homozygotes from 3,222 healthy British Pakistani adults, with high levels of consanguinity, including only variants with minor allele frequency below 1% (Narasimhan *et al.* 2016). Only mutations affecting 100% of transcripts were taken forward to the analysis, to avoid mutations affecting only minor splicing variants. This identified 36 IMPI genes with at least one LoF homozygote.
- A dataset of LoF homozygotes from 10,503 Pakistani adults, including only variants with minor allele frequency below 1% (Saleheen *et al.* 2017). Only those marked as ‘confident pLoF’ (predicted loss-of-function) were taken forward. This identified 73 IMPI 217 genes with at least one LoF homozygote.
- Data from the ExAC (Exome Aggregation Consortium) resource (Lek *et al.* 2016). LoF homozygotes/hemizygotes were identified with below 1% minor allele frequency and no annotation/confidence flags, which were predicted to affect at least the canonical transcript. This identified 122 IMPI genes with at least one LoF homozygote.

Isoforms were identified from Ensembl transcripts associated with each gene (Yates *et al.* 2016), with proteins retrieved from the corresponding UniProt ID (The UniProt Consortium 2017). Only protein-encoding transcripts not labelled as fragments were included.

Statistics

Statistical analyses were carried out in R. Statistical tests utilised and any correction for multiple testing are indicated in the text.

Results

Phylogenetic spread of monogenetic disease genes of the mitochondrion

Genes associated with monogenetic disease in humans have a broader phylogenetic profile (López-Bigas & Ouzounis 2004; Kondrashov & Koonin 2004) and are more likely to originate in the early Metazoa (Domazet-Lošo & Tautz 2008), than those not associated with monogenetic disease. I first asked whether the phylogenetic differences between disease and non-disease genes observed in analyses of the total human genome are replicated in the subset of the genes encoding the mitochondrial proteome, using the previously described dataset of IMPI 2017 orthologues (*Chapter 2*).

Initially, I compared the distributions of the number of orthologues predicted for genes with associated monogenetic disease to those without any association with disease (*Figure 5.1*). Neither distribution was normally distributed (Shapiro-Wilk test; non-disease, $p < 2.2 \times 10^{-16}$; disease, $p = 7.3 \times 10^{-10}$), so I used nonparametric tests to test differences between the distributions. The distributions of numbers of predicted orthologues of non-disease and disease genes were significantly different (Mann-Whitney U test; $p < 2.2 \times 10^{-16}$). Genes not associated with disease had a lower number of predicted orthologues, on average – a mean of 109 orthologues compared to a mean of 153 orthologues for known disease genes. This difference held when looking at genes in the IMPI positive training set (Mann-Whitney U test; $p = 8.56 \times 10^{-12}$), looking at only the number of eukaryotic orthologues (Mann-Whitney U test; $p < 2.2 \times 10^{-16}$) and looking at only the number of orthologues from holozoans (Mann-Whitney U test; $p < 2.2 \times 10^{-16}$).

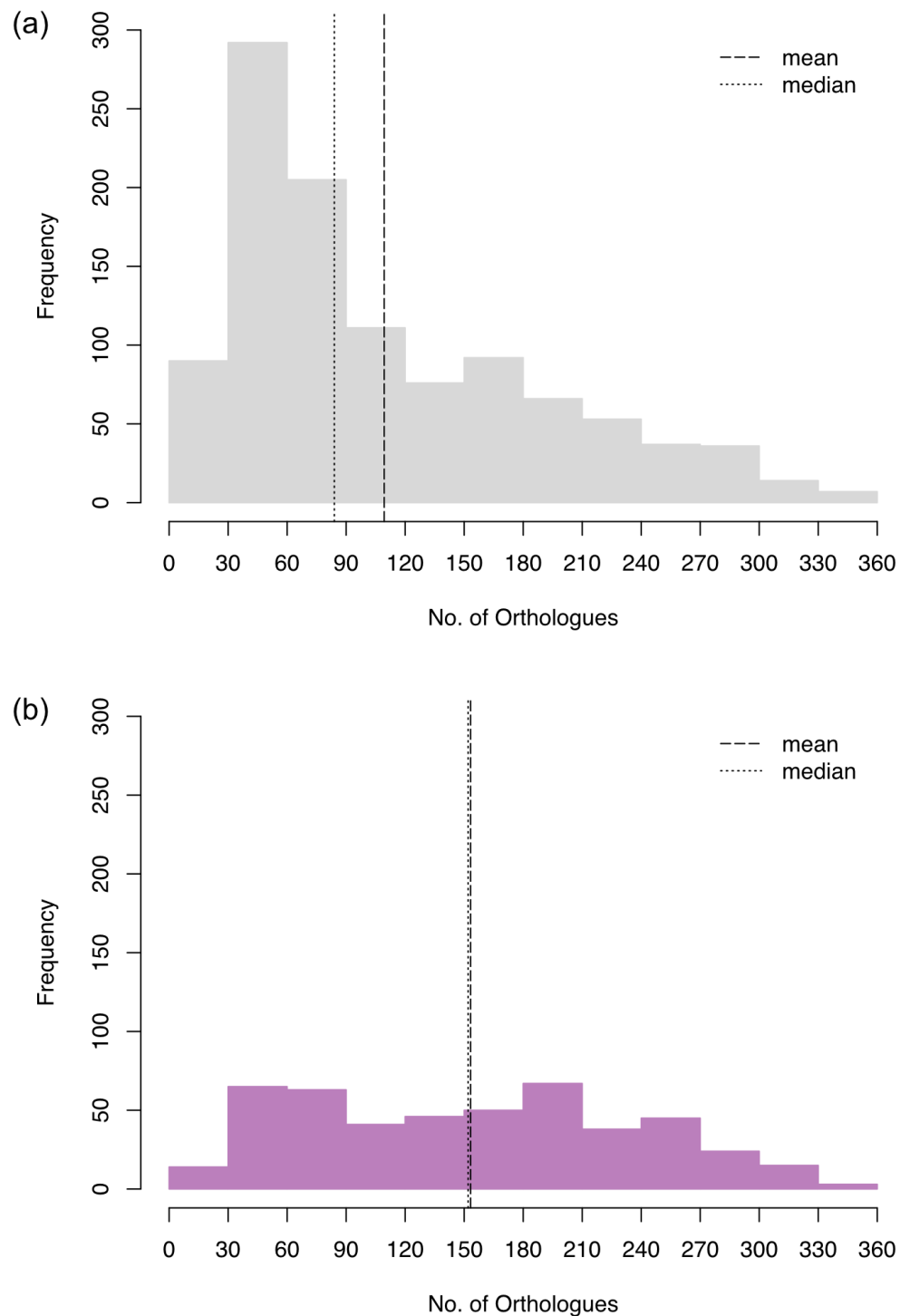


Figure 5.1 IMPI genes not associated with disease have fewer predicted orthologues. (a) Distribution of IMPI genes not associated with monogenetic disease. (b) Distribution of IMPI genes associated with monogenetic disease. Distributions are significantly different (Mann-Whitney U test, $p < 2.2 \times 10^{-16}$).

The distributions of orthologues will be influenced by the number of species of different types included in the analysis – for example, if a large number of fungi are included compared to the number of amoeba, the orthologue counts will be influenced more by whether there are fungal orthologues than amoebal orthologues. Therefore, I then looked at the spread of orthologues across eight defined taxa: Holozoa, Holomycota, Amoebozoa, Archaeplastida, SAR group, Excavata, Archaea and Bacteria (*Figure 5.2*). There was a significant difference between the number of taxa with predicted orthologues for IMPI disease and non-disease genes (χ^2 test; $p < 2.2 \times 10^{-16}$). Nearly half (47%) of the genes not associated with monogenetic disease were predicted in only one taxon, whereas only 25% of disease genes showed the same result. Thus, IMPI disease genes are more likely to have a wider spread throughout the eight studied taxa than IMPI genes not associated with disease.

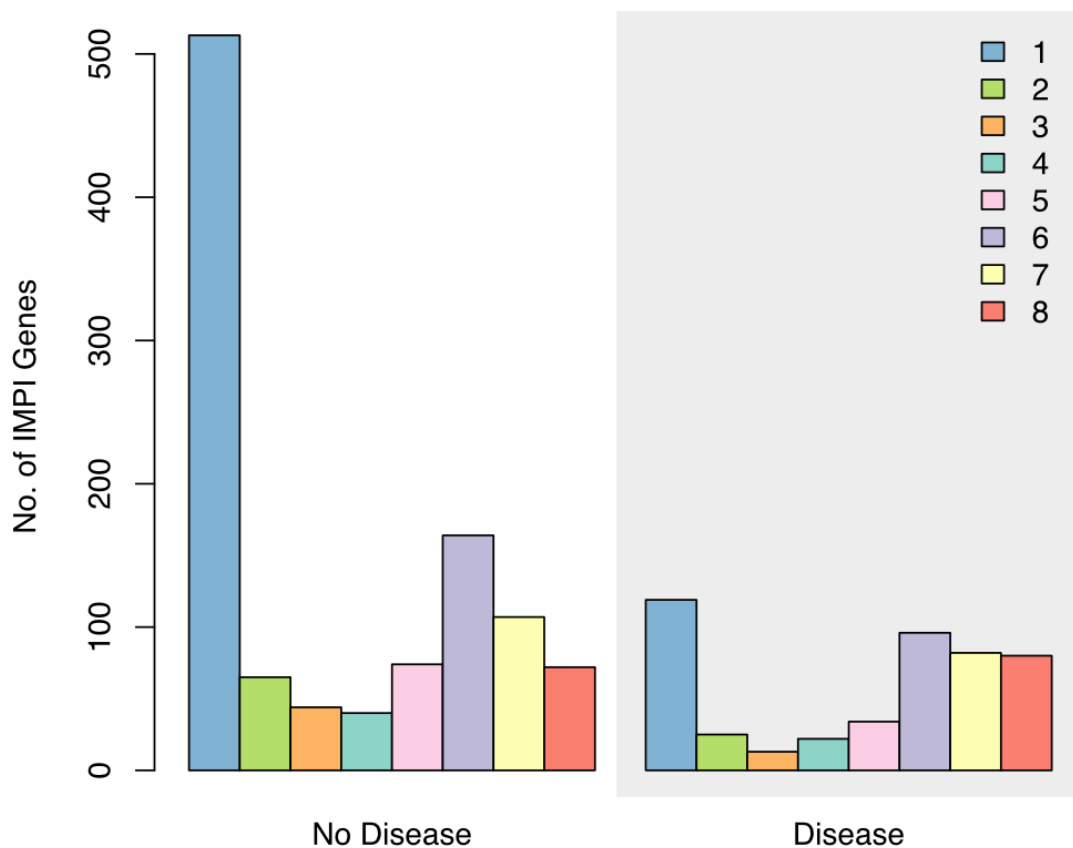


Figure 5.2 Differences between the number of taxa with predicted orthologues for disease and non-disease genes (χ^2 test, $p < 2.2 \times 10^{-16}$). Horizontal scale is the number of taxa for which orthologues of each gene are identified, from: Holozoa, Holomycota, Amoebozoa, Archaeplastida, SAR group, Excavata, Archaea and Bacteria.

Phylogenetic origin of monogenetic disease genes of the mitochondrion

Another way to look at the data is to consider the group in which orthologues are first identified. In general, the oldest genes are present in prokaryotes, whereas the newest genes will be present only in holozoan eukaryotes. I assigned each gene to one of the groups Holozoa, non-holozoan eukaryota or Prokaryota, based on the first appearance of a predicted orthologue (*Table 2.2*). There was a significant difference in the phylogenetic origin of IMPI genes associated with disease and IMPI genes with no associated disease (χ^2 test; $p < 2.2 \times 10^{-16}$), replicated within the IMPI positive training genes (χ^2 test; $p = 1.1 \times 10^{-14}$). Though disease genes did originate in all three of the designated groups, 45% of genes with earliest orthologues predicted in prokaryotes were associated with monogenetic disease, compared to only 19% of genes with orthologues predicted only in holozoans.

Table 5.2 *Earliest taxa with predicted orthologues of IMPI genes, separated by known monogenetic disease-causing status.*

	Earliest orthologue identified in:		
	Holozoa	Non-holozoan Eukaryota	Prokaryota
No disease	513 (81%)	344 (67%)	222 (54%)
Disease	119 (19%)	167 (33%)	185 (45%)

Phylostratigraphic analysis is a more detailed way of investigating the origin of genes (Domazet-Lošo & Tautz 2008). Genes are assigned to a phylostratum – a key taxonomic division in evolutionary history, such as the evolution of the vertebrates – based on the most phylogenetically distant prediction of an orthologue. The phylostrata are linearly separated, based around the species under investigation (in this case, humans). This can be used to identify a phylostratum from which different human genes originate. In this case, I separated the origin of genes into fourteen phylostrata – key divisions in the evolutionary history of humans (*Figure 5.3*).

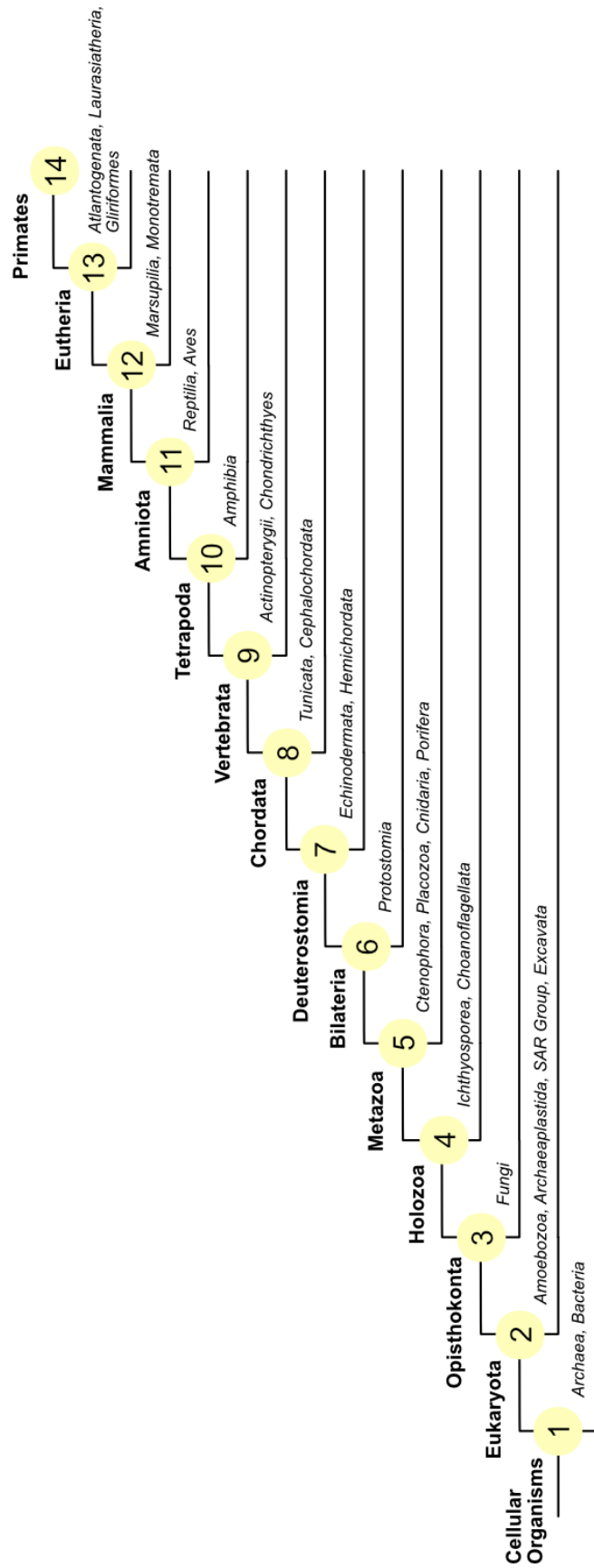


Figure 5.3 Fourteen phylostrata used in phylostratigraphic analysis. Points of separation used in further analysis are numbered. Each phylostratum is labelled with the additional taxa defining the separation.

IMPI genes showed three main peaks of phylostratic origin (*Figure 5.4*). The earliest was at the two phylostrata ‘Cellular organisms’ (1) and ‘Eukaryota’ (2) – genes which trace back to the origin of life with cellular organisms and the origin of eukaryotes. A slightly larger proportion of IMPI genes were assigned to the phylostratum ‘Eukaryota’ (2), compared to the total human genome, where a larger proportion were assigned to the phylostratum ‘Cellular organisms’ (1) (Domazet-Lošo & Tautz 2008). This may reflect gene innovations supporting the function of the new mitochondrion in eukaryotes. For example, twenty-two IMPI genes assigned to ‘Eukaryota’ (2) were designated the GO Biological Process term ‘transport’ (GO:0006810), but none of the genes assigned to ‘Cellular organisms’ (1) were designated this term. This can be explained as transport across a mitochondrial membrane is unnecessary in single-compartment organisms.

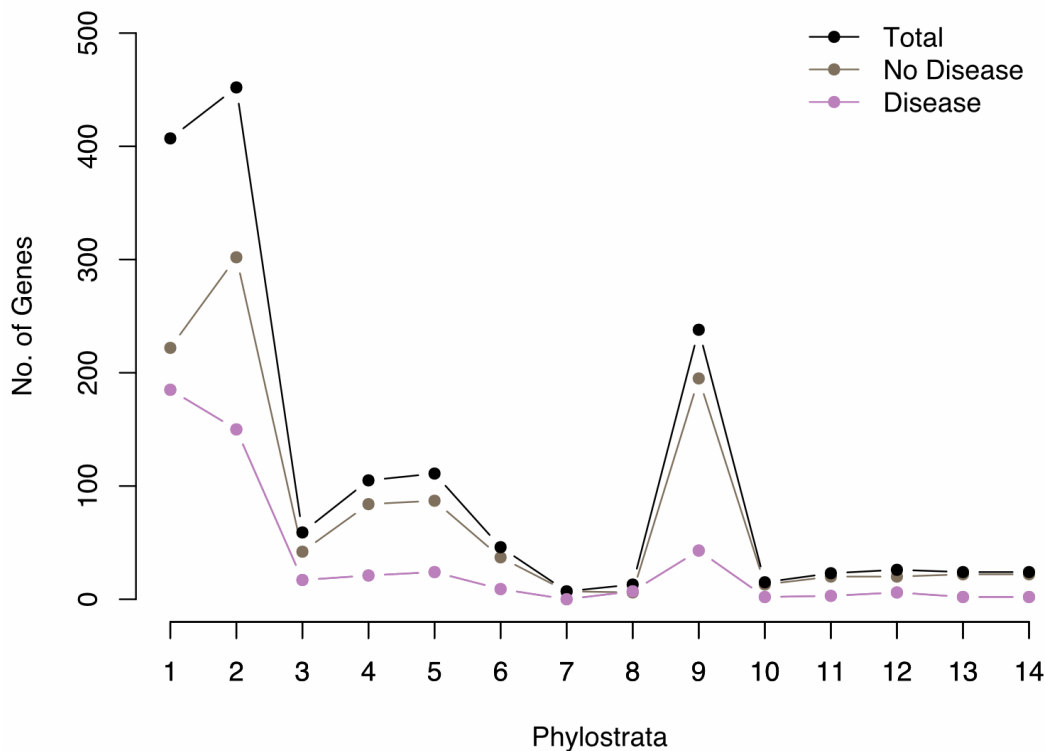


Figure 5.4 Phylostratigraphic analysis of all IMPI genes (black; N = 1,550), IMPI 2017 genes not associated with disease (grey; N = 1,079), and IMPI 2017 genes associated with monogenetic disease (purple; N = 471). Phylostrata are numbered as described in Figure 5.3.

There was a second smaller peak at the phylostrata ‘Holozoa’ (4) and ‘Metazoa’ (5), which includes the origin of animals and their closest unicellular relatives, before the Cambrian explosion. The third and clearer peak was at the phylostratum for the origin of the vertebrates (9), also noted in the analysis of the full human genome where the peak is associated with

immunity-related genes (Domazet-Lošo & Tautz 2008). No GO Biological Process term was overrepresented in IMPI genes originating in vertebrates compared to the rest of IMPI 2017. However, the largest number of genes were assigned terms ‘apoptotic process’ (GO:0006915; 19 genes) and ‘signal transduction’ (GO:0007165; 17 genes).

Figure 5.4 suggests there are small differences between the phylostratic origins of IMPI disease and non-disease genes, particularly around the phylostratic divisions at ‘Cellular organisms’ (1) and ‘Eukaryota’ (2). To assess this difference, I plotted the log odds ratio between disease and non-disease genes for each investigated phylostratum (Figure 5.5).

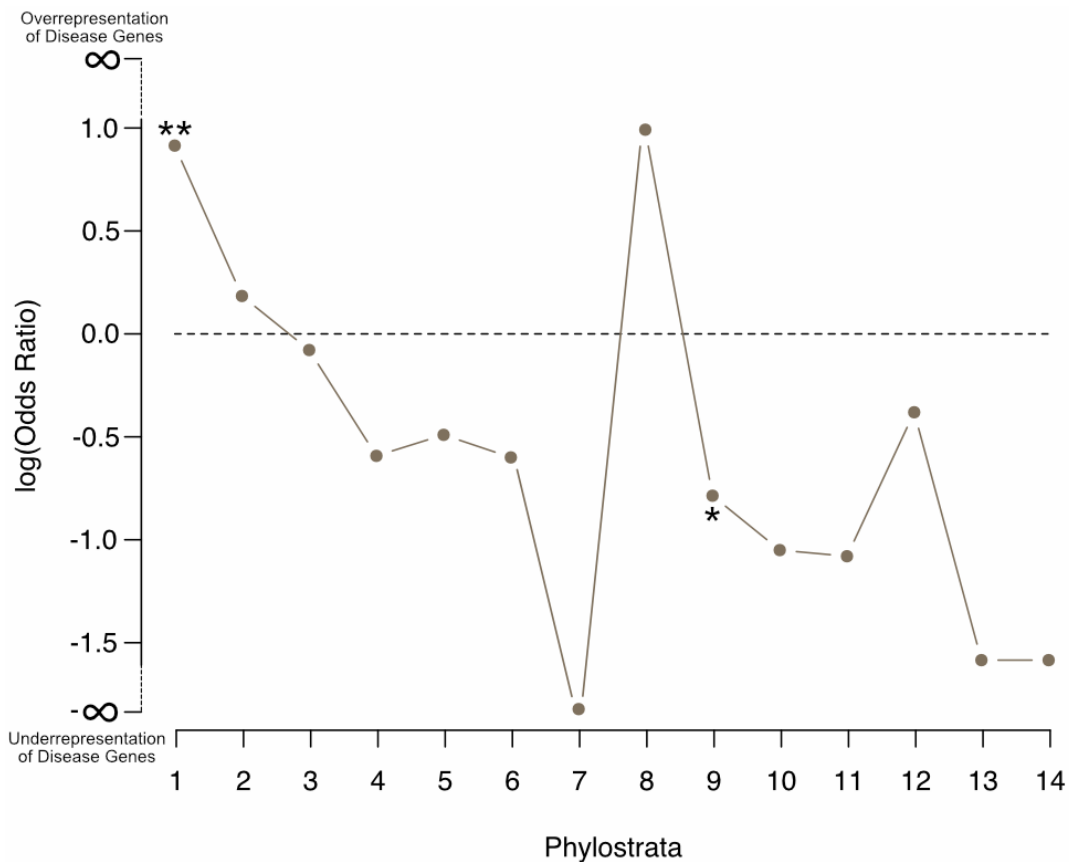


Figure 5.5 Comparing the representation of disease and non-disease IMPI genes in different phylostrata. Natural log odds ratio is plotted for each phylostratum. Significance tested for each phylostratum using Fisher’s exact test with Bonferroni correction for multiple testing (* $p = 5.36 \times 10^{-5}$; ** $p = 7.19 \times 10^{-13}$). Phylostrata are numbered as described in Figure 5.3.

Two phylostrata showed significant differences for the log odds ratio between disease and non-disease genes. IMPI genes with a phylostratic origin in the cellular organisms (1), before the evolution of eukaryotes, were significantly more likely to be associated with disease, with a non-logged odds ratio of 2.5 (Fisher’s exact test; $p = 7.2 \times 10^{-13}$). Disease genes originating

in cellular organisms were enriched for basic metabolic processes, represented by one cluster of GO Biological Process terms with an enrichment score of 1.44 (*Table 5.3*), though no terms were statistically enriched individually.

Table 5.3 Enriched cluster of GO Biological Process terms for disease-causing genes originating in cellular organisms, in comparison to all IMPI genes. Enrichment score = 1.44.

GO Biological Process term	Count	<i>p</i> -value	Benjamini corrected <i>p</i> -value
Tricarboxylic acid cycle	11	7.9×10^{-4}	0.18
Pyruvate metabolic process	6	0.01	0.60
Acetyl-CoA biosynthetic process from pyruvate	4	0.02	0.69
Mitochondrial acetyl-CoA biosynthetic process from pyruvate	3	0.17	0.99
Regulation of acetyl-CoA biosynthetic process from pyruvate	4	0.19	0.99
Glucose metabolic process	4	0.41	1

However, IMPI genes with a phylostratic origin in the vertebrates were significantly less likely to be associated with disease, with a non-logged odds ratio of 0.46 (Fisher's exact test; $p = 5.4 \times 10^{-5}$). No individual GO Biological Process term was enriched for the non-disease genes originating in vertebrates, with the largest cluster related to apoptosis with an enrichment score of 1.11 (*Table 5.4*).

Table 5.4 Enriched cluster of GO Biological Process terms for non-disease genes originating in vertebrates, in comparison to all IMPI genes. Enrichment score = 1.11.

GO Biological Process term	Count	<i>p</i> -value	Benjamini corrected <i>p</i> -value
Positive regulation of protein insertion into mitochondrial membrane involved in apoptotic signalling pathway	4	0.04	0.99
Cellular response to hypoxia	6	0.06	0.99
Regulation of mitochondrial membrane permeability	4	0.08	1
Positive regulation of intrinsic apoptotic signalling pathway	4	0.11	1
Release of cytochrome <i>c</i> from mitochondria	4	0.14	1

Inheritance patterns of IMPI genes associated with monogenetic mitochondrial disease

Monogenetic diseases can show different patterns of inheritance. Previous work had identified some differences between human genes associated with dominant disease (only one faulty allele causes disease) and genes associated with recessive disease (two faulty alleles necessary to cause disease) (Kondrashov & Koonin 2004; Cai *et al.* 2009), but no-one had explored their phylogeny beyond the yeast *S. cerevisiae*, or investigated genes of the mitochondrial proteome in particular. I explored disease genes in IMPI split by mode of inheritance, with the hypothesis that there is a difference in phylogenetic spread and origin between dominant and recessive IMPI disease genes.

First, I compared the distributions of numbers of orthologues of dominant disease genes and recessive disease genes from the previously described orthologue dataset (*Chapter 2*). I removed from the analysis any gene associated with diseases of both recessive and dominant inheritance or genes from the sex chromosomes, as these will be under different types of selective pressure. Neither distribution was normally distributed (Shapiro-Wilk test; recessive disease genes, $p = 4.5 \times 10^{-7}$; dominant disease genes, $p = 9.1 \times 10^{-6}$), so nonparametric statistics were used. There was a significant difference between the distribution of orthologue numbers between dominant disease genes and recessive disease genes in IMPI (*Figure 5.6*; Mann-Whitney *U* test; $p = 0.0009$). On average, dominant disease genes had fewer orthologues than recessive disease genes (median of 122 orthologues compared to 168 orthologues). This difference was significant when comparing only IMPI positive training set genes (Mann-Whitney *U* test; $p = 2.8 \times 10^{-5}$), only comparing numbers of eukaryotic orthologues (Mann-Whitney *U* test; $p = 0.01$), and even comparing numbers of holozoan orthologues (Mann-Whitney *U* test; $p = 0.004$) despite the low numbers of dominant disease genes in the dataset (only 55 in total).

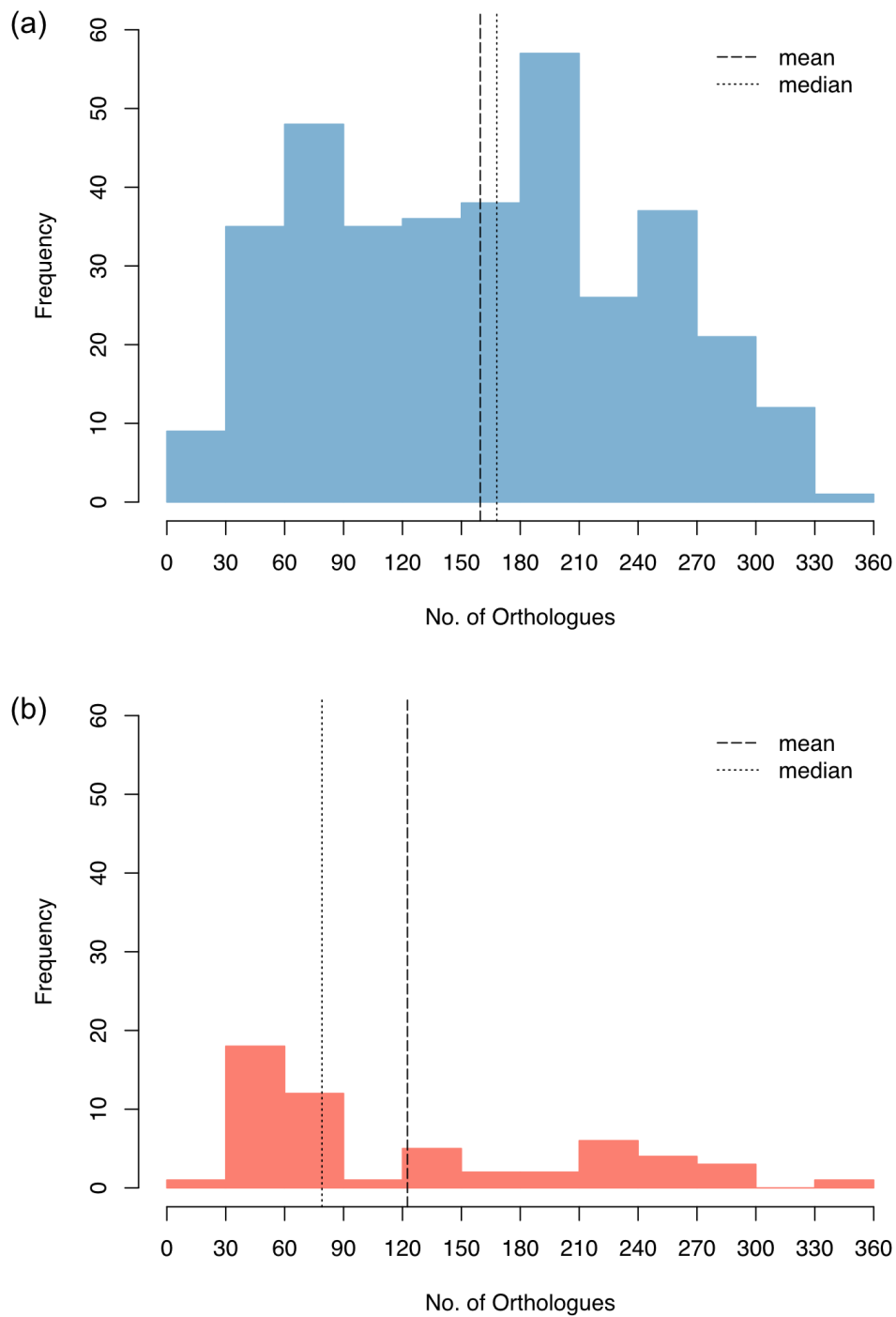


Figure 5.6 IMPI genes associated with dominant diseases have fewer orthologues on average. (a) Distribution of IMPI genes associated with recessive disease. (b) Distribution of IMPI 2017 genes associated with dominant disease. Distributions are significantly different (Mann-Whitney U test, $p = 0.0009$).

Next, I explored the spread of the disease gene orthologues through the eight previously defined taxa: Holozoa, Holomycota, Amoebozoa, Archaeplastida, SAR Group, Excavata, Archaea and Bacteria (Figure 5.7). There was a significant difference between the spread of predicted

orthologues of recessive and dominant IMPI disease genes throughout these taxa (χ^2 test; $p = 2.2 \times 10^{-5}$). The majority (54%) of dominant disease genes only had predicted orthologues in one taxon, whereas nearly half (49%) of recessive disease genes had predicted orthologues in between six and eight taxa. Overall, the spread of recessive disease genes over different taxa was wider than the spread of dominant disease genes.

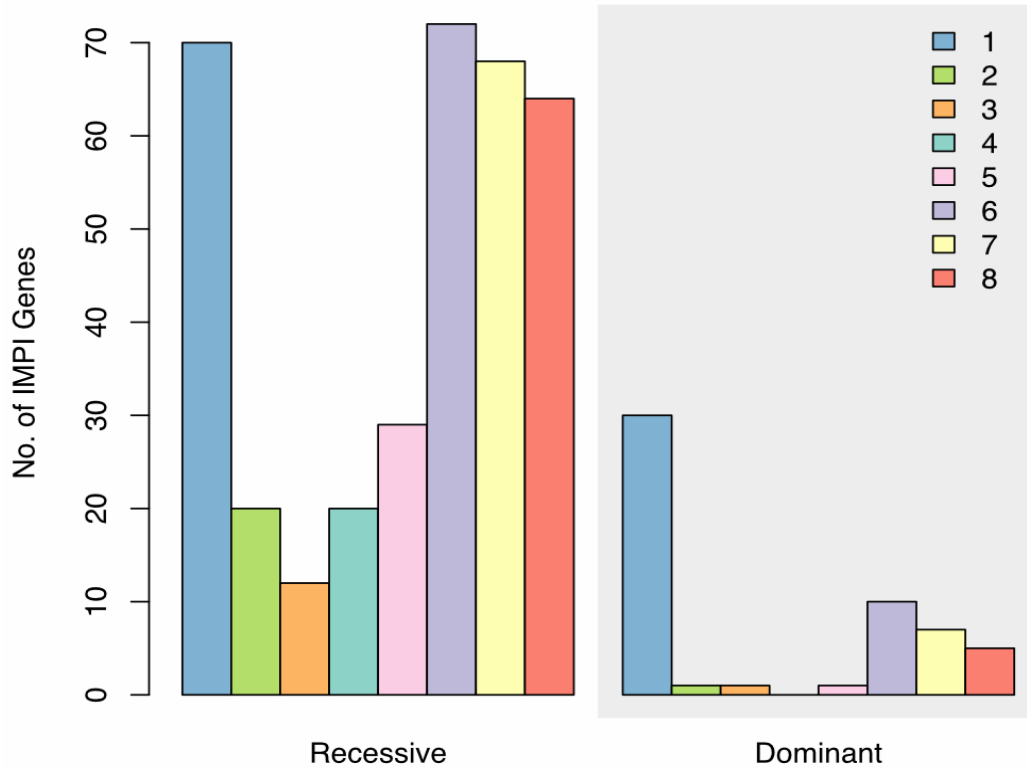


Figure 5.7 Differences between the number of taxa with predicted orthologues of recessive and dominant IMPI disease genes (χ^2 test, $p = 1.3 \times 10^{-5}$). Horizontal axis shows the number of taxa with orthologues of each gene in: Holozoa, Holomycota, Amoebozoa, Archaeplastida, SAR group, Excavata, Archaea and Bacteria.

I then investigated the origin of these mitochondrial disease genes, by dividing the dominant and recessive disease genes into three groups dependent on where the first orthologue was predicted: Holozoa, non-holozoan eukaryota and Prokaryota (Table 5.5). There was a significant difference in the origin of dominant and recessive IMPI disease genes (χ^2 test; $p = 1.5 \times 10^{-7}$), which was consistent within the IMPI positive training set (χ^2 test; $p = 5.2 \times 10^{-9}$). Over half of all the dominant disease genes in IMPI were novel to the Holozoa ($30/55 = 54\%$), whereas 80% of recessive disease genes had an older origin.

Table 5.5 Earliest taxa with predicted orthologues of IMPI disease-associated genes, separated by known inheritance type.

	Earliest orthologue identified in:		
	Holozoa	Non-holozoan eukaryota	Prokaryota
Dominant	30 (30%)	11 (7%)	14 (9%)
Recessive	70 (70%)	136 (93%)	149 (91%)

As before, I used phylostratigraphic analysis to investigate the origin of recessive and dominant disease genes in more detail (*Figure 5.8*). Mitochondrial disease genes showed the same general pattern as total IMPI genes, with the two largest peaks of innovation at the phylostrata defining the separation of cellular life/eukaryotes (1/2); and at the phylostratic origin of vertebrates (9). This was clear even when looking at only the dominant disease genes in the analysis, although there were comparatively few of them – only 55 compared to 355 recessive disease genes. 76% of recessive disease genes had early phylostratic origins, before the separation of the opisthokonts (between 904 and 1,579 million years ago (Eme *et al.* 2014)), compared to only 45% of dominant disease genes.

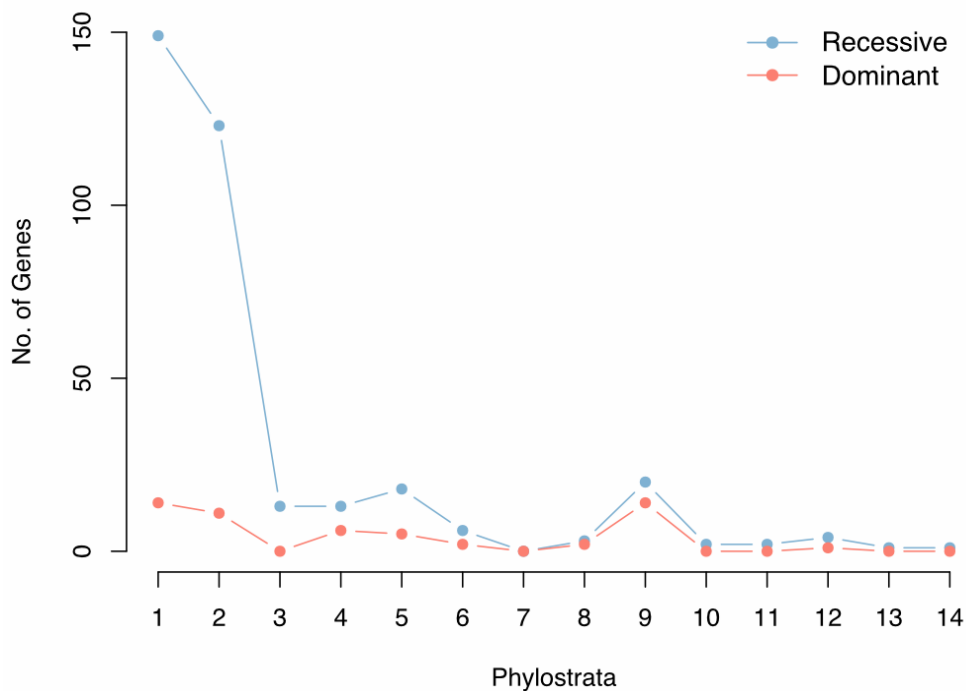


Figure 5.8 Phylostratigraphic analysis of IMPI disease genes, separated into recessive (blue; N = 355); and dominant (red; N = 55). Phylostrata are numbered as described in Figure 5.3.

However, the differences between recessive and dominant disease genes in these early phylostrata were not significant (*Figure 5.9*) – the p -value for genes assigned to the phylostratum ‘Cellular organisms’ (1) was 0.36 and for ‘Eukaryota’ was 0.44, using Fisher’s exact test.

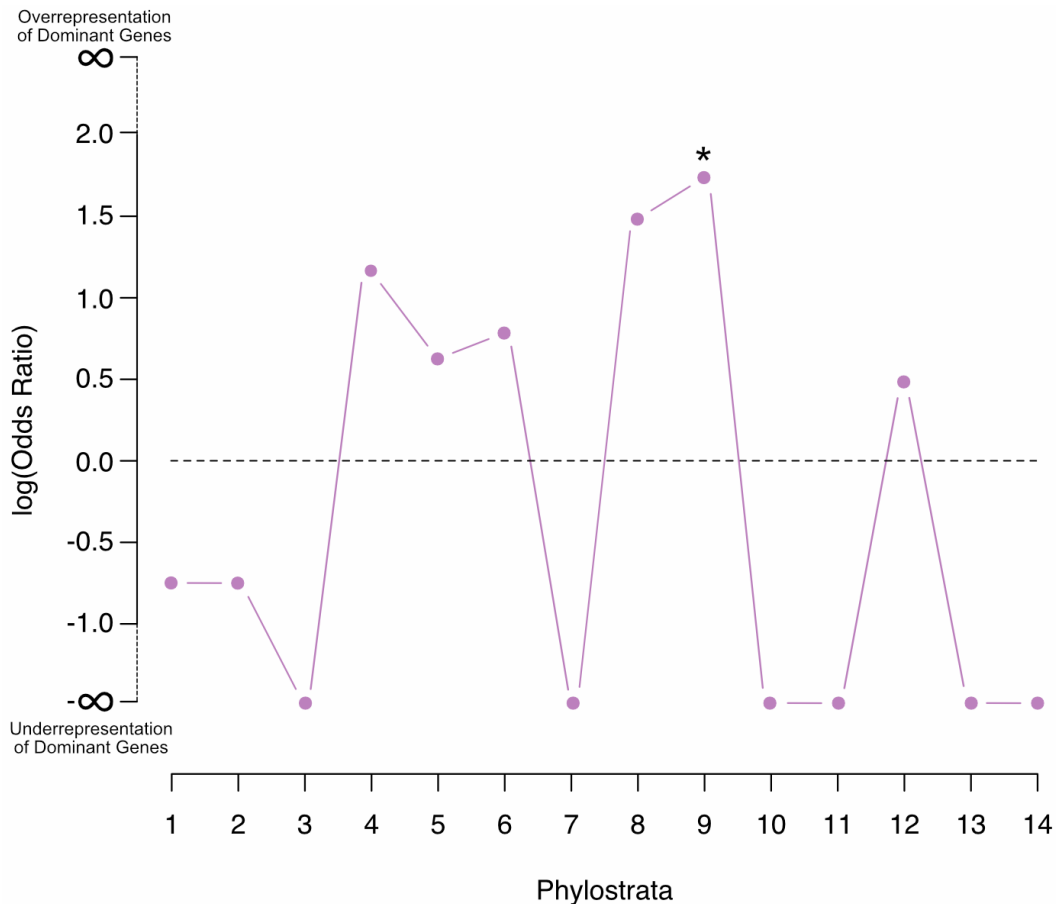


Figure 5.9 Comparing representation of dominant and recessive IMPI disease-associated genes in different phylostrata. Natural log odds ratio is plotted for each phylostratum. Significance tested for each phylostratum using Fisher’s exact test with Bonferroni correction for multiple testing (* $p = 0.0003$). Phylostrata are numbered as described in Figure 3.

The only phylostratum with a significant difference between the proportions of dominant disease genes and recessive disease genes was the ‘vertebrata’ (9) (*Figure 5.9*). Disease genes with described dominant inheritance were significantly overrepresented in the disease genes originating within the vertebrates, compared to those with recessive inheritance, (Fisher’s exact test; $p = 0.0003$) with a non-logged odds ratio of 5.7. No GO Biological process term was individually significantly overrepresented in dominant disease genes assigned to the vertebrate phylostratum. However, two functional clusters were significant (*Table 5.6*). These

clusters both contained genes important in apoptosis and signalling, as well as control of cell proliferation, with strong overlap between the genes in the two clusters.

Table 5.6 Enriched clusters of GO Biological Process terms for dominant disease-causing genes originating in vertebrates, compared to the genes forming the total IMPI set.

GO Biological Process term	Count	<i>p</i> -value	Benjamini corrected <i>p</i> -value
<i>Cluster 1 (Enrichment score: 2.69)</i>			
Negative regulation of cysteine-type endopeptidase activity involved in apoptotic process	4	1.4×10^{-4}	0.053
Signal transduction	5	4.3×10^{-4}	0.082
Negative regulation of apoptotic process	5	9.8×10^{-4}	0.12
Positive regulation of transcription from RNA polymerase II promoter	4	1.7×10^{-3}	0.15
Cell proliferation	4	1.8×10^{-3}	0.12
Positive regulation of peptidyl-serine phosphorylation	3	2.2×10^{-3}	0.12
Platelet activation	3	5.1×10^{-3}	0.23
Intracellular signal transduction	3	0.038	0.73
<i>Cluster 2 (Enrichment score: 2.39)</i>			
Negative regulation of apoptotic process	5	9.8×10^{-4}	0.12
Negative regulation of neuron death	3	1.7×10^{-3}	0.13
Aging	3	0.041	0.72

Functional analysis

Another way to investigate the characteristics of disease genes is to look at the functions of the proteins encoded by these genes. Previous work showed genes with enzyme activity (GO:0003824 ‘catalytic activity’) were more likely associated with recessive disease than dominant disease (Kondrashov & Koonin 2004). However, there was no significant difference between dominant and recessive IMPI disease genes, when looking at association with the this GO term (*Figure 5.10a*; χ^2 test; $p = 0.13$). The percentage of dominant disease genes with the annotation was slightly lower than recessive disease genes (62% compared to 73%). There was also no significant difference between the association with metabolism (GO:0008152 ‘metabolic process’) between dominant and recessive IMPI disease genes (*Figure 5.10b*; χ^2 test; $p = 0.16$), though again dominant disease genes were slightly less likely to be annotated with this term than recessive disease genes (75% compared to 83%).

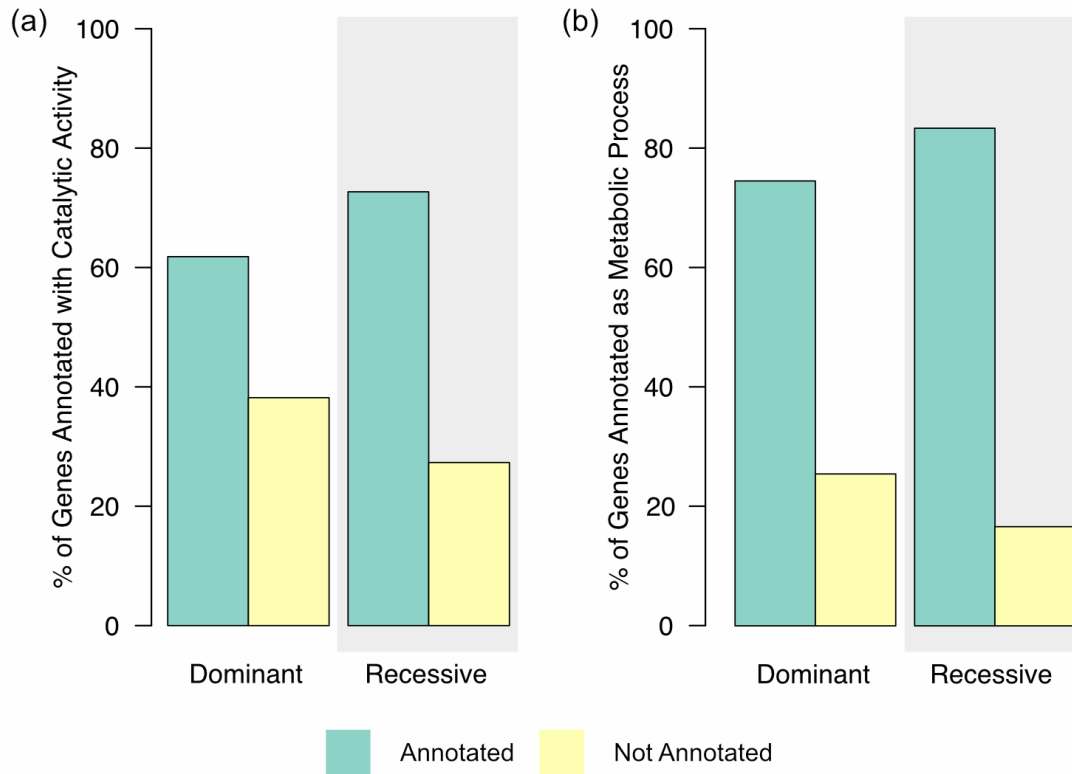


Figure 5.10 Comparing the percentage of IMPI dominant and recessive mitochondrial disease genes with certain GO annotations. (a) Annotation ‘catalytic activity’ (GO:0003824). No significant difference (χ^2 test, $p = 0.13$). (b) Annotation ‘metabolic process’ (GO:0008152). No significant difference (χ^2 test, $p = 0.16$).

These GO terms are both very broad. 866 (56%) IMPI genes were annotated with the GO term ‘catalytic activity’ and 1,111 (72%) IMPI genes were annotated with the GO term ‘metabolic process’. I, therefore, tried a more restrictive and specific comparison, comparing genes catalysing reactions in central mitochondrial metabolism with those that were not, as defined by inclusion in the MitoCore model of central metabolism (Smith *et al.* 2017) (Figure 5.11). There was a significant difference between the proportions of dominant and recessive IMPI disease genes involved in central mitochondrial metabolism (χ^2 test; $p = 0.0016$). IMPI genes associated with recessive disease were more likely involved in central mitochondrial metabolism than those associated with dominant disease.

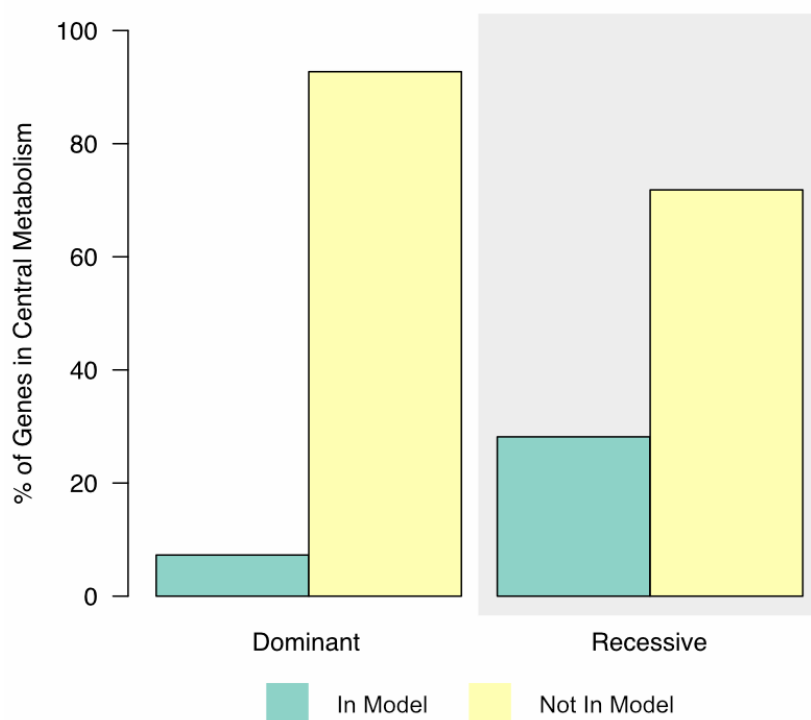


Figure 5.11 Percentage of IMPI dominant and recessive disease genes in the MitoCore mitochondrial model of central metabolism. There is a significant difference between the annotation status of the two types of disease inheritance (χ^2 test, $p = 0.0016$).

There was also a significant difference between the evolutionary origin of IMPI genes and their involvement in central mitochondrial metabolism (Figure 5.12; χ^2 test; $p = 8.4 \times 10^{-6}$). Genes which originated more recently were less likely to be involved in central mitochondrial metabolism than more ancient genes – 37% of genes originating in prokaryotes compared to 21% of genes originating in non-holozoan eukaryotes and 12% of genes originating in holozoan eukaryotes. This may suggest a possible link between function, origin of the disease and gene inheritance.

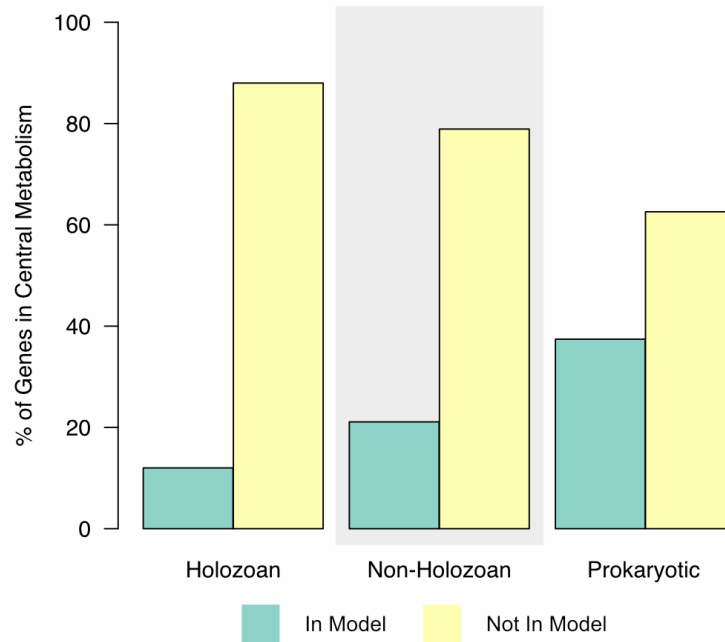


Figure 5.12 Percentage of IMPI genes of different origin in the MitoCore model of central metabolism. There is a significant difference between the annotation status of the genes originating in the different groups (χ^2 test, $p = 8.4 \times 10^{-6}$).

Human loss-of-function (LoF) homozygotes

Though I have classified genes as ‘disease’ or ‘non-disease’, the ‘non-disease’ genes will include genes that cause currently unassigned genetic diseases, as well as genes that, when mutated, potentially cause problems so large that the foetus is inviable. I decided to create a cleaner set of genes negatively associated with monogenetic disease for further analysis, using datasets of human loss-of-function (LoF) homozygotes – people with two copies of mutations predicted to destroy all gene function, but who had no symptomatic phenotype. These genes may also be useful in prioritisation of candidate disease genes, as genes with healthy human LoF homozygotes have been shown to have non-deleterious effects, and so may be moved down in terms of prioritisation of disease gene candidates.

In total, 236 IMPI genes were identified with loss-of-function mutations affecting both alleles of at least one study participant, from three published studies on consanguineous populations (Sulem *et al.* 2015; Narasimhan *et al.* 2016; Saleheen *et al.* 2017) and the ExAC dataset, which contains information on 60,706 unrelated individuals from a variety of ethnic

backgrounds (Lek *et al.* 2016). Despite the strict criteria used for identification of LoF homozygotes across the data sources, 35 (15%) of these genes are associated with human monogenetic disease. Most of these have reasonable explanations for the observation of at least one seemingly healthy LoF human homozygote (*Table 5.7*). These genes were removed from the LoF set.

Table 5.7 Possible explanations for human LoF homozygotes in IMPI genes with associated monogenetic disease(s). Genes can be assigned to multiple groups.

Possible Explanation	Genes
Unaffected isoforms	<i>ATP7B, CARS2, COQ2, ELAC2, FANCG, FBXO7, GDAP1, NDUF51, OPA3, WARS2, TPP1, YME1L1</i>
Mild/variable described phenotype	<i>ABCB6, DMGDH, IVD, PHKA1, PRIMPOL, PRODH, SLC25A13</i>
Late onset/undiagnosed	<i>FANCG, GDAP1, PHKA1, SPG7</i>
Change in function	<i>NOL3</i>
Unknown	<i>PDHA1, TAZ</i>

For twelve of the genes, the LoF mutation does not affect all isoforms and the unaffected isoforms may provide enough function to protect from disease. For example, the *OPA3* gene encodes two nearly identical proteins produced by alternative splicing (*Figure 5.13*) – the 179 aa isoform skips exon 3, whilst the 180 aa isoform skips exon 2. All currently described disease-causing mutations affect either both the 179 aa and 180 aa isoforms or just the canonical 179 aa isoform, whilst the LoF homozygote affects only the 180 aa isoform. This suggests the function of the 180 aa isoform is not essential for human health. A third shorter isoform is not described in the literature.

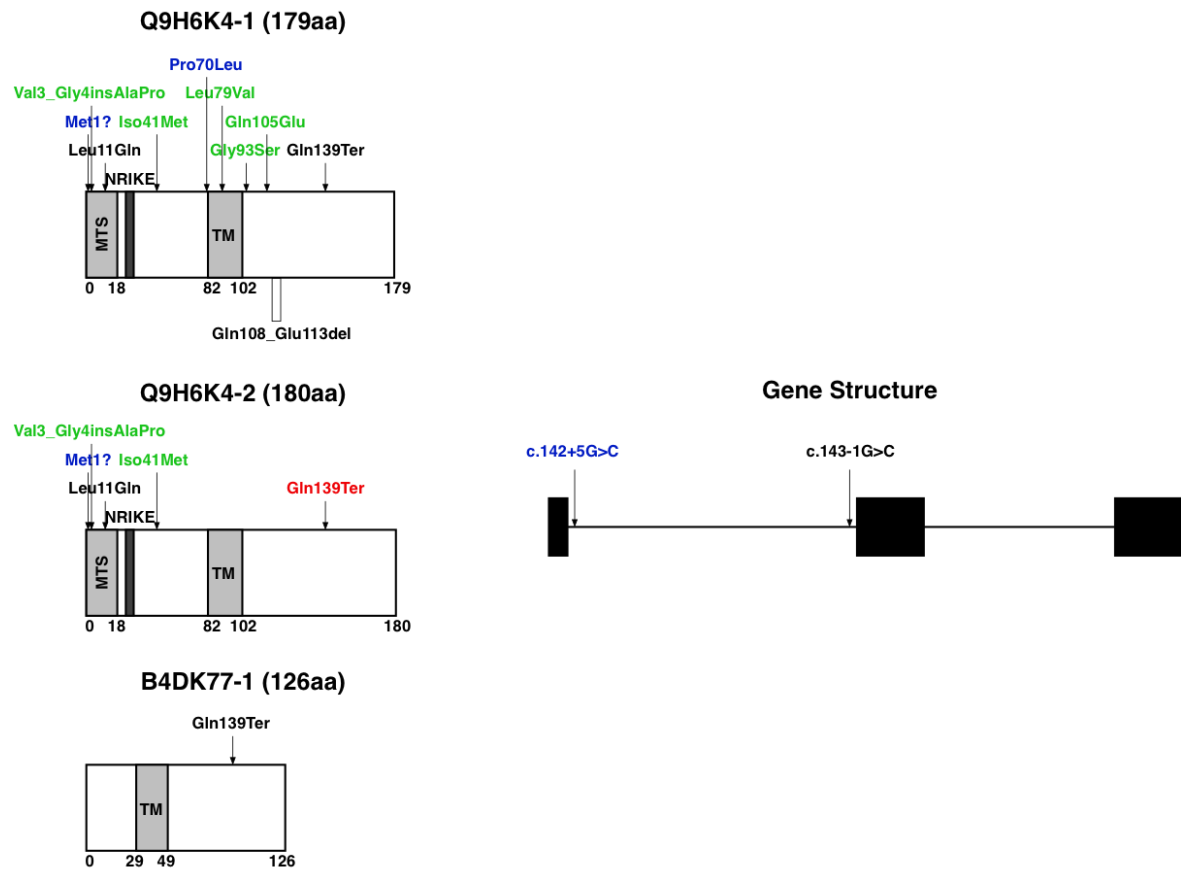


Figure 5.13 OPA3 gene and protein isoform structures, annotated with known disease-causing mutations. Red = LoF homozygote, green = heterozygous disease mutation, blue = compound heterozygous disease mutation, black = homozygous disease mutation. The LoF mutation is only located on the slightly longer 180 aa isoform. (MTS = predicted mitochondrial targeting sequence; TM = transmembrane domain; NRIKE = internal mitochondrial sorting signal).

Other genes are associated with late onset disease so, though the study participant may seem healthy, they may be affected later in life. Four of the genes possibly fit this criterion (*Table 5.7*). For example, the homozygous p.Leu78Ter LoF mutation in *SPG7* identified in a single subject from a Pakistani cohort (Saleheen *et al.* 2017) has been described in spastic paraplegia patients in homozygosity (Arnoldi *et al.* 2008).

Equally, seven of the disease genes have described diseases with very mild effects or have patients with very variable phenotypes, suggesting variable effects or incorrect assignment of pathogenicity (*Table 5.7*). For example, a heterozygous mutation in *PRIMPOL* has been described as causative of severe myopia in a Chinese family (Zhao *et al.* 2013) and functional studies showed decreased primase and polymerase activity (Keen *et al.* 2014). However, a literature search showed further work has identified the mutation in controls at a similar

frequency to patients (insignificant χ^2 test) (Li & Zhang 2015). The identification of this LoF homozygote gives further evidence to the idea that decreased function of *PRIMPOL* does not cause severe myopia.

This analysis showed that while healthy human LoF homozygotes may be a useful tool in prioritising candidate disease genes, it is important to look in depth at the results of the mutation (e.g. looking at isoforms) and to consider the type of disease (e.g. LoF homozygotes may not be useful in studies of late-onset disease).

I carried out gene enrichment analysis to investigate whether any KEGG pathways or GO Biological Function terms were over-represented in the mitochondrial gene LoF homozygotes. No pathway or biological function was over-represented (no p -value < 0.05). One cluster of functional terms had an enrichment score above one (*Table 5.8*), representing processes involving fatty acids.

Table 5.8 Enriched cluster of GO Biological Process and KEGG pathway terms for human LoF homozygotes within IMPI. Enrichment score = 1.55.

Annotation type	Annotation	Count	p -value	Benjamini corrected p -value
GO Biological Process	Fatty acid biosynthetic process	9	0.0046	0.97
GO Biological Process	Acetyl-CoA metabolic process	7	0.02	0.99
KEGG Pathway	Butanoate metabolism	5	0.23	0.88

Phylostratigraphic analysis showed there was no significant difference in the phylostratic origin of IMPI genes with healthy human LoF homozygotes and other IMPI genes without associated disease (*Figure 5.14*). There was a significant negative correlation between the numbers of the phylostrata and log odds ratio, assuming each phylostratum is equally separated from the next (Pearson's product-moment correlation; $r = -0.77$, $p = 0.001$). This suggested that older non-disease genes were less likely to have identified healthy LoF homozygotes.

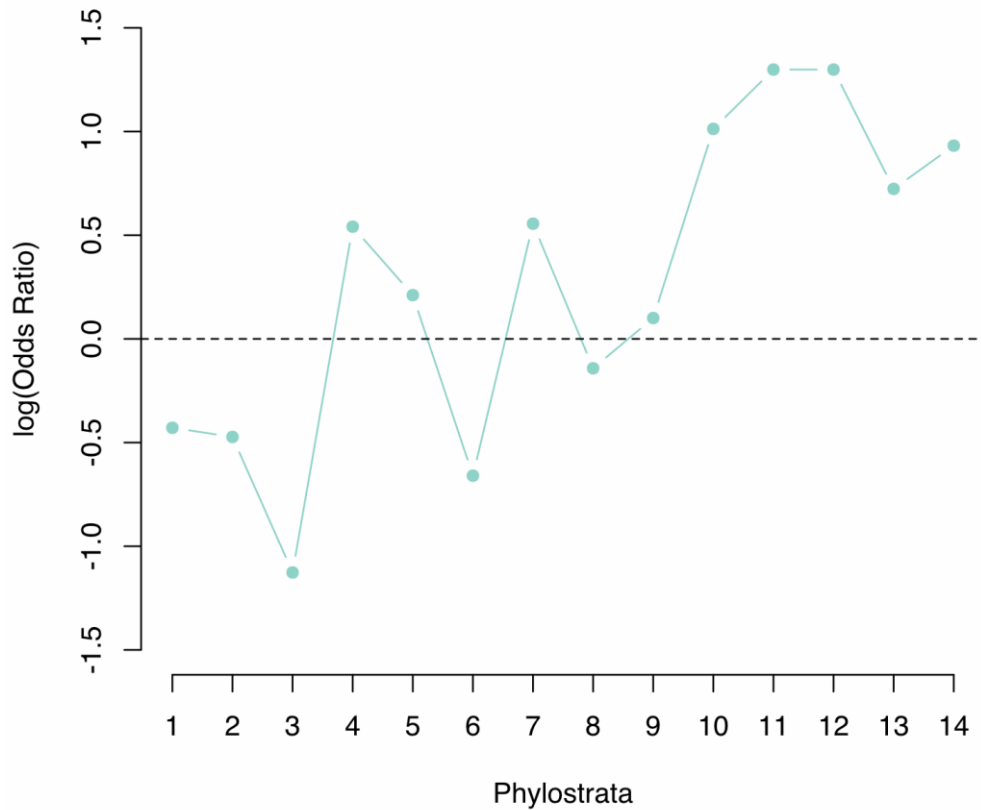


Figure 5.14 Comparing the origin of IMPI genes with LoF homozygotes and non-disease IMPI genes in different phylostrata. Natural log odds ratio is plotted for each phylostratum. Significance tested for each phylostratum using Fisher's exact test with Bonferroni correction for multiple testing (no significant phylostrata). Phylostrata are numbered as in Figure 5.3.

Essential genes

Studies of human genes and their association with disease often rely on orthologous genes in model organisms. Three of the most popular model organisms are mice (*Mus musculus*), baker's yeast (*Saccharomyces cerevisiae*) and the γ -proteobacterium *Escherichia coli*. These organisms have either full or partial gene knockout collections with attached phenotype data. Although previous work suggested genes essential to mouse did not strongly reflect 'essential' genes in humans (Liao & Zhang 2008), I investigated whether the essentiality of genes in several model organisms could be used to prioritise candidate genes for different types of human mitochondrial disease.

Although knockout strains of nearly all yeast and *E. coli* genes have been created, this is not true for mice. Therefore, analysis of mouse orthologues of human IMPI genes was restricted to the 800 genes with a descriptive mouse knockout phenotype. Yeast and *E. coli* analyses included all genes with a predicted IMPI orthologue. Where at least one non-symptomatic human LoF homozygote and a disease were both associated with the same gene, the gene was grouped with the disease genes.

There was a significant difference between the known human disease status of essential and non-essential genes in mouse, yeast and *E. coli* (Table 5.9; χ^2 test; $p = 0.007$, $p = 0.03$ and $p = 0.008$, respectively). Disease genes in human were more likely to be essential in mouse and yeast (6% and 20% of disease genes were essential, respectively) than genes not currently associated with disease and genes with human LoF homozygotes (3% and 17% of genes not currently associated with disease, respectively); but less likely to be essential in *E. coli* (20% of disease genes were essential compared to 37% of non-disease genes). These differences, while significant overall, were small. Human IMPI genes with non-symptomatic human LoF homozygotes were very unlikely to be essential in mouse (0%) and yeast (0%), but slightly more likely to be essential in *E. coli* (22.2%).

Table 5.9 IMPI genes with knockout models in three model organisms (mouse, *S. cerevisiae* and *E. coli*) separated by essentiality in the model organism and disease status in humans. Genes which are both associated with a disease and a predicted human LoF homozygote were grouped with the disease genes.

	Disease	No disease	Human LoF
<i>Mouse</i>			
Essential	19	15	0
Non-essential	279	409	123
<i>S. cerevisiae</i>			
Essential	39	47	0
Non-essential	155	200	28
<i>E. coli</i>			
Essential	22	45	4
Non-essential	86	71	14

Very few of the essential genes in any of three model organisms had described LoF homozygotes in humans, as would be expected with a basic level of functional conservation. Two exceptions in the eukaryotic model organisms come from genes which had both a monogenetic disease association and a predicted human LoF homozygote – these genes were

grouped in with the disease-associated genes in *Table 5.9*. *PDHA1* (pyruvate dehydrogenase E1 alpha subunit) is an essential gene in mouse which is also associated with a homozygous LoF human variant. A single hemizygous male is recorded in ExAC, with a C → G mutation causing a p.Tyr49Ter truncation of the canonical transcript. *PDHA1* changes are recorded as the cause of an X-linked dominant disease in humans – pyruvate dehydrogenase E1-alpha deficiency (OMIM:300502), including a patient with a premature termination at amino acid 313 (Matthews *et al.* 1994), a later point than recorded in this male. There are a variety of phenotypes associated with *PDHA1* deficiency, but all are early-onset. Analysis of splicing variants also did not provide an obvious explanation for this male.

There was also one essential gene in yeast with an identified human LoF homozygote: *TRZ1* – the orthologue of human *ELAC2* (ElaC ribonuclease Z 2). A single human LoF homozygote has been described in a person from an African population, sampled in ExAC, with a splice acceptor variant (rs149733287). Though the mouse orthologue is marked non-essential, the phenotype of the mouse knockout is ‘preweaning lethality, complete penetrance’. It seems unlikely, therefore, that the predicted human LoF mutation causes complete loss-of-function. There are several alternatively spliced transcripts marked as protein coding in Ensembl, which may compensate for the splice site change.

Knockout models reflect only one specific cause of a phenotype change – that caused by a complete loss of gene function. This is most likely to model the cause of a recessive disease correctly, where function of both alleles has been sufficiently changed or lost to cause a phenotype. However, dominant diseases are caused by a loss or gain in function of only one allele, and so are not well represented by complete knockouts. I investigated whether essential genes were more likely to be causative of dominant diseases in the three studied model organisms (*Table 5.10*). There was a significant difference between the mode of human disease inheritance and the essentiality of a mouse or yeast knockout for IMPI disease genes (Fisher’s exact test; $p = 1.0 \times 10^{-8}$ and $p = 0.02$, respectively), but not for essentiality of *E. coli* knockouts (Fisher’s exact test; $p = 0.27$). Essential genes in mouse were more likely to be associated with dominant disease (87%), but those in yeast (23%) and *E. coli* (9%) were not. This may relate to the (at least partial) haploid lifestyles of *E. coli* and yeast.

Table 5.10 Human IMPI genes with knockout mouse models and associated human monogenetic disease, divided by essentiality in mouse and the mode of disease inheritance in humans.

	Recessive	Dominant
<i>Mouse</i>		
Essential	2	13
Non-essential	187	33
<i>S. cerevisiae</i>		
Essential	23	7
Non-essential	126	11
<i>E. coli</i>		
Essential	17	2
Non-essential	70	3

The two essential mouse genes associated solely with recessive disease are *SLC25A26* (a mitochondrial transporter for *S*-adenosylmethionine) and *SUCLA2* (a subunit of succinyl-CoA synthetase). Patients with *SLC25A26* mutations have varied residual activity of the transporter, with the severity of the symptoms correlating with the residual transport activity (Kishita *et al.* 2015). Not all function has been lost as in the mouse knockout model, so it may be that complete loss-of-function is lethal in humans as in mouse. Though a variety of *SUCLA2* deficient patients have been described, I could not find direct measurements of succinyl-CoA synthetase activity in these patients.

The association of mouse essential genes with dominant disease in humans may help prioritise candidates for dominantly inherited disease in genetically uncharacterised human patients. To support this idea, I looked at a measurement associated with both dominant disease and haploinsufficiency (inability of one functional copy of a gene to sustain sufficient function) in humans – the pLI (probability of being loss-of-function intolerant) score (Lek *et al.* 2016).

As in the total human genome (Lek *et al.* 2016), IMPI dominant disease genes had a significantly higher average pLI score (Figure 5.15) than recessive disease genes (Welch's *t*-test; $p = 2.8 \times 10^{-11}$) and non-disease genes (Welch's *t*-test; $p = 4.3 \times 10^{-8}$). I also compared the pLI scores of genes with non-symptomatic human LoF homozygotes to the other three groups, finding a significant difference with non-disease genes and with dominant disease genes (Welch's *t*-test; $p = 3.1 \times 10^{-15}$ and $p = 7.4 \times 10^{-13}$ respectively). This is consistent with using the pLI score as an indicator of possible disease genes, particularly for dominant

diseases. The boxplot for ‘non-disease’ genes shows the likely heterogeneous nature of this group – 11% of these genes had a pLI score above 0.8 compared to 2% of genes with LoF human heterozygotes. These high scoring pLI genes are good candidates for possible dominant disease genes in humans, though they may also reflect genes so essential to human function that any change is lethal. This second category may never be observed in human disease patients, and study of the essentiality of these genes will rely on model organisms.

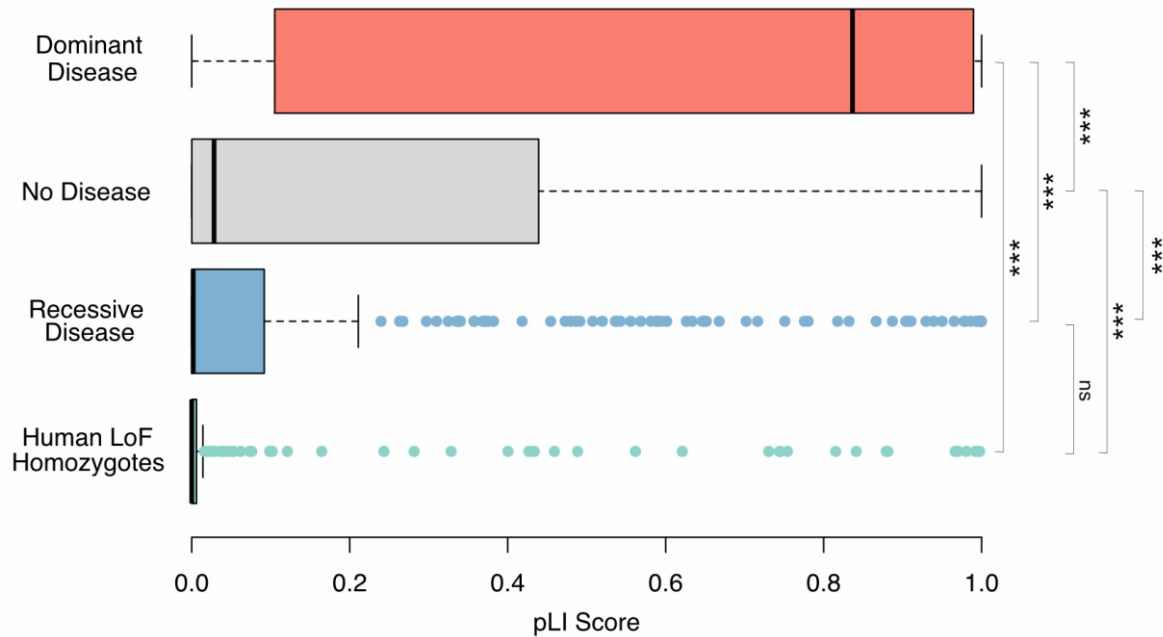


Figure 5.15 Boxplot of pLI Scores for IMPI genes. Comparing genes with human LoF homozygotes ($N = 200$), recessive disease genes ($N = 353$), dominant disease genes ($N = 55$) and ‘no disease’ genes ($N = 855$). p-values from Welch’s t-tests, with Bonferroni correction for multiple testing (*** $p < 0.001$; ns = not significant).

As a step towards identifying these essential genes, I then looked at only the genes essential to mouse which are currently not associated with disease. There was a significant difference between the pLI scores of these genes and the pLI scores of non-essential mouse genes not currently associated with disease (Figure 5.16; Welch’s t-test; $p = 0.03$). Essential genes in mouse not currently associated with disease had a higher mean pLI score (0.51) than non-essential genes (0.23).

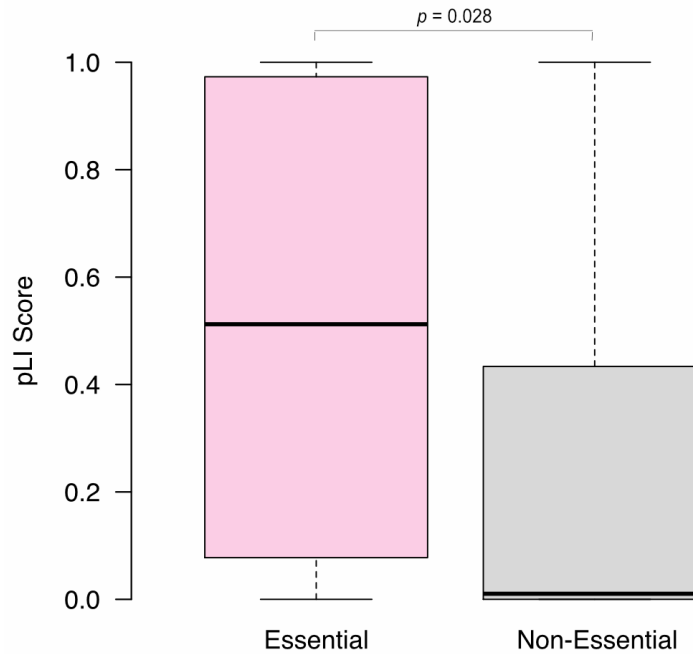


Figure 5.16 Boxplot of pLI scores of human IMPI genes not currently associated with disease or a LoF homozygote, separated by essentiality of the orthologue in mice. Comparing genes which are essential ($N = 15$) in mouse to those which are non-essential ($N = 522$). p-value from Welch's t-test.

Six genes had a pLI score over 0.9 (Table 5.11), which may be the best candidates for genes essential in both humans and mice, or dominant disease genes in humans.

Table 5.11 IMPI genes not currently associated with a monogenetic disease in humans, which are essential when knocked-out in mice and have a pLI score of over 0.9.

Ensembl ID	Gene name	Gene description	pLI score
ENSG00000147162	<i>OGT</i>	O-Linked N-acetylglucosamine Transferase	1.00
ENSG00000149428	<i>HYOU1</i>	Hypoxia Up-regulated 1	1.00
ENSG00000067225	<i>PKM</i>	Pyruvate Kinase, Muscle	0.99
ENSG00000100030	<i>MAPK1</i>	Mitogen-Activated Protein Kinase 1	0.98
ENSG00000108953	<i>YWHAE</i>	Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Epsilon	0.96
ENSG00000167004	<i>PDIA3</i>	Protein Disulphide Isomerase Family A Member 3	0.96

Discussion

Previous work had investigated some phylogenetic and evolutionary features of human disease genes. Mitochondria have a unique evolutionary history, as descendants of an endosymbiotic α -proteobacterium (Sagan 1967; Müller & Martin 1999). Therefore, I used a dataset of orthologues of members of the mitochondrial proteome (as defined by IMPI) to explore the phylogenetic and evolutionary features of mitochondrial disease genes.

Known monogenetic disease genes in IMPI have, on average, a wider spread of orthologues throughout the tree of life, and originate earlier than, IMPI genes without associated disease (*Figures 5.1 – 5.5, Table 5.2*). This is consistent with previous analyses which had looked at genome-wide analyses of human monogenetic disease genes (López-Bigas & Ouzounis 2004; Domazet-Lošo & Tautz 2008; Maxwell *et al.* 2014), despite the unique history of the mitochondria. There are two large peaks of mitochondrial gene innovation: one in the cellular organisms and early eukaryotes – these are enriched with disease-causing genes; and the second at the origin of the vertebrates – these are enriched with genes not associated with disease. The genes originating in the vertebrates include a large number which are associated with apoptosis and signalling, which is consistent with a rise in complexity of the immune system in vertebrates (Aravind *et al.* 2001) and an increase in the complexity of development controlled by additional signalling (Pires-da Silva & Sommer 2003). This is different than when looking at the same comparison genome-wide, where a greater majority of genes are of newer origin and other phylostrata show significant differences (e.g. genes originating in eukaryotes are depleted in disease-causing genes, whereas there is no significant difference in the mitochondrial analysis) (Domazet-Lošo & Tautz 2008). However, that study looked more at gene families than individual genes, as the authors do not try and separate the origin of paralogues within the same family. The analysis in this chapter may be more useful in specifically prioritising candidate genes for mitochondrial disease patients.

In my reading of the literature I could not find any previous work which had looked in depth at the phylogenetic spread and origin of disease genes showing different types of inheritance patterns – namely recessive and dominant. IMPI genes associated with dominant disease have, on average, a narrower phylogenetic spread and are of more recent origin than IMPI genes associated with recessive disease (*Figures 5.6 – 5.9, Table 5.5*). A previous study found a similar but qualitative difference in the origin of recessive and dominant disease genes

across the complete human genome, but only looked at orthologues in the Metazoa and *S. cerevisiae*, and recorded no significance testing (Cai *et al.* 2009).

The difference in origin of dominant and recessive IMPI disease genes can be, at least partially linked to function. Recessive disease genes are more likely associated with central metabolism in mitochondria (*Figure 5.11*), as are older genes (*Figure 5.12*). Dominant genes are more likely to originate in vertebrates (*Figure 5.9*) and are enriched in processes such as cell proliferation and apoptosis (*Table 5.6*). This is consistent with a previous hypothesis which suggested that increased complexity in Metazoa necessitated expansion of a variety of protein types, including those involved in transcription, signalling and cellular structures (Jimenez-Sanchez *et al.* 2001).

However, there was no significant link between enzyme function as defined by the GO term ‘catalytic activity’ (GO:0003824) and inheritance type of the disease gene (*Figure 5.10*), as seen previously (Kondrashov & Koonin 2004). This may be due to the increase in annotation assigning catalytic activity to a wider range of genes since the original analysis, or a real difference between a genome-wide analysis and an investigation of only mitochondrial genes.

Though the previously mentioned work was based on splitting IMPI into disease-causing genes and non-disease-causing genes, the non-disease group is likely a heterogeneous mix, containing genes which really are non-essential to human function, disease-causing genes which have yet to be identified, and genes which are so essential to function we may never identify human patients with functional mutations. I explored known human LoF homozygotes of IMPI genes, identified from a variety of sources, as a set of genes which are likely to be truly non-essential and which should perhaps be deprioritised as candidate disease genes. These genes showed no significant difference in phylostratic origin compared to other IMPI genes not associated with disease (*Figure 5.14*). However, examples of human LoF homozygote genes that are associated with disease highlight several considerations to using LoF homozygotes to rule genes out for further analysis in patients (*Table 5.10*) – for example, does the LoF mutation affect all splice variants and was the patient’s disease late-onset?

The closest we may get to identifying essential genes is by studying model organisms. However, previous work had suggested there was not a strong correlation between essentiality in mouse and in human (Liao & Zhang 2008). Essential genes in mouse and yeast were slightly but significantly more likely to be associated with disease (*Table 5.10*), indicating

that human patients are born and diagnosed with mutations in genes which are essential in the model organisms. However, most of the diseases in human associated with essential genes from mouse were dominantly inherited (*Table 5.11*). These genes also had a significantly higher pLI score (*Figure 5.16*), suggesting essential genes in mouse are most likely associated with either dominant disease or haploinsufficiency in humans. This suggests knockout model organisms, particularly mouse, could be used in prioritising gene candidates for dominant disease or for identifying the most essential genes in humans.

Even with better prioritisation of candidate genes and an increase in success rate, exome sequencing will not provide sufficient data to identify the cause for all genetic mitochondrial disease. Other causes may be identified, for example, by investigating copy number variants (CNVs) – larger scale deletions and/or duplications – which are difficult to identify from exome data (Krumm *et al.* 2012). Whole genome sequencing allows better detection of structural changes, as well as analysis of non-exonic sequences. A recent study showed that whole genome sequencing improved diagnostic rates in paediatric patients with suspected genetic disease (Lionel *et al.* 2018). Mitochondrial disease has also been associated with digenic gene inheritance (Van Goethem *et al.* 2003) – the contribution of mutations in two separate, though related, genes to a single phenotype, which may be considered in cases which are proving intractable to solve.

Conclusions

In this chapter, I have investigated the origin and phylogenetic spread of mitochondrial proteome-encoding genes associated with monogenetic disease. Disease associated genes are more likely to have a wider spread of orthologues throughout the tree of life than those not associated with disease, though genes associated with dominant disease have a narrower spread and more recent origin than genes associated with recessive disease. Disease genes are also more likely to have originated earlier in human evolutionary history than genes not associated with disease. Genes associated with dominant disease, however, are more likely to have originated in the Holozoa than recessive disease genes, and this may be linked to the overrepresentation of ancient genes of central metabolism in the recessive disease gene group. This information could be used to prioritise candidate genes from exome sequencing of mitochondrial disease patients. Study of model organisms with gene knockouts and healthy human LoF homozygotes may inform candidate gene prioritisation for patients with different types of inherited mitochondrial disease.

Chapter 6

Mitochondrial proteins in viruses

Introduction

Viruses and the mitochondria

Mitochondria are involved in a variety of cellular processes which could be of value to a virus, in terms of survival and replication in a eukaryotic cell. These include:

- energy metabolism, including the production of ATP via the electron transport chain, which may be manipulated to favour viral replication,
- apoptosis (controlled cell death), which may be important in cellular immune responses against viruses, and in the escape of replicated viruses from a host cell (Galluzzi *et al.* 2008),
- nucleotide and amino acid metabolism, which could provide source materials for viral replication,
- DNA replication, as the host mitochondrial machinery could be hijacked by the virus,
- and protein translation, for the same reason (mitochondria contain their own ribosome and associated machinery).

There are many examples where viruses have been shown to manipulate some of these mitochondrial processes (Ohta & Nishiyama 2011). For example, human cytomegalovirus infection is associated with changes in several mitochondrial associated processes. Energy metabolism is manipulated to increase flux through glycolysis for energy (Munger *et al.* 2006) and to increase flux through the tricarboxylic acid cycle via glutamine metabolism (Chambers *et al.* 2010). Fatty acid synthesis is also increased to support production of the viral envelope (Munger *et al.* 2008). Infection with this virus also induces mitochondrial biogenesis, with an associated increase in mitochondrial transcription and translation (Karniely *et al.* 2016). These are just a few examples, from a single virus, of the many ways mitochondria can be manipulated.

Specific virally-encoded proteins have been linked to changes in mitochondrial function on infection with some viruses. For example, the UL12 gene in herpes simplex virus type 1 encodes an exonuclease that causes the degradation of mitochondrial DNA (Corcoran *et al.* 2009), which has been linked to mitochondrial damage as an important part of early viral infection (Wnęk *et al.* 2016).

Viral orthologues of mitochondrial proteins

Viral genomes can encode proteins which are orthologous with mitochondrial proteins from eukaryotes. Some of these orthologous proteins are mitochondrially localised in humans or other eukaryotes, though the virally encoded proteins may also localise to other subcellular compartments (Ohta & Nishiyama 2011). For example, there are several identified viral orthologues of the Bcl-2 family of proteins across different viral families (Cuconati & White 2002), which function in the control of the mitochondrial pathway of apoptosis in humans (Gross *et al.* 1999). The first identified viral member of the Bcl-2 family was the E1B-19K protein encoded in the genome of an adenovirus (White *et al.* 1991) – a group of viruses known to cause illness in a range of vertebrates. E1B-19K is an anti-apoptotic member of the family, preventing the breakdown of DNA (White *et al.* 1991). A more recent study of E1B-19K has shown that the modification of the host cell death process produced by this protein reduces the response of the host innate immune system, which may increase the success of viral replication (Radke *et al.* 2014). There are both pro-apoptotic and anti-apoptotic members of this family across the viral proteome, as there are in humans. One example of a pro-apoptotic viral Bcl-2 family protein is protein 7A encoded by the RNA virus, ‘severe acute respiratory syndrome corona virus’. This protein is associated with an increase in the efficiency of apoptosis but is not necessary for viral replication (Schaecher *et al.* 2007).

Whilst viral orthologues of several mitochondrial proteins have been identified, there has not been a large-scale study of this phenomenon or the general types of proteins involved. The development of a new set of genes predicted to encode the human mitochondrial proteome (IMPI 2017) provides the opportunity to make this assessment, facilitating insight into the importance of controlling or manipulating the host cell mitochondria across known viruses.

Chapter summary

In this chapter, I explore the viral proteome for viral protein orthologues of the human mitochondrial proteome. I investigate the types of viruses with and without predicted orthologues; the numbers of orthologues identified within different viruses; and the families of genes with predicted viral orthologues. I explore one example of possible horizontal

transfer of a gene encoding a mitochondrial protein to a virus from its host cell. I then investigate the predicted viral orthologues of the mitochondrial carrier family.

Methods

Identifying viral orthologues of human mitochondrial genes

A database of viral protein sequences was created by downloading all viral sequences (NCBI taxid: 10239) from the NCBI protein database (www.ncbi.nlm.nih.gov/protein). For all genes in IMPI 2017 (www.mrc-mbu.cam.ac.uk/impi) and MitoCarta 2.0 (Calvo *et al.* 2015) (two different definitions of the human mitochondrial proteome), the canonical protein sequence in humans was identified. This was used in a reciprocal best hit analysis to identify possible viral orthologues (*Figure 6.1*), using an E-value threshold of 10 for each BLASTp search. The human sequence was first used as the bait for a BLASTp search against the database of all viral sequences. For each viral strain with a hit, the top scoring hit was used in a BLASTp search back against a database of human protein sequences. If the top hit from this search matched a protein encoded by the original human gene used to search the viral sequence database, a viral orthologue was predicted.

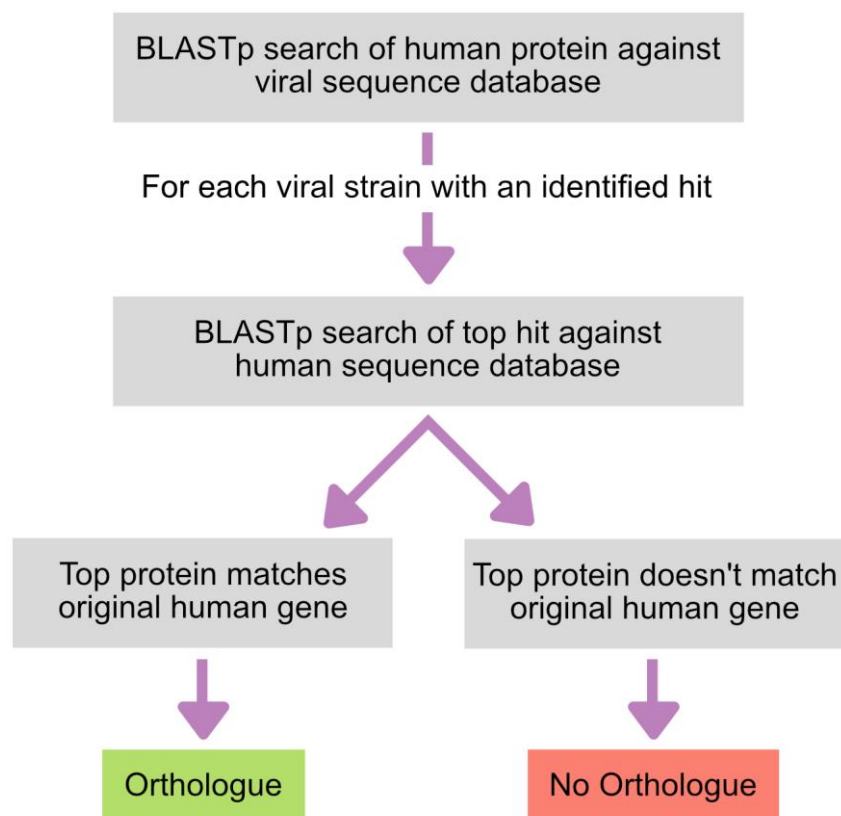


Figure 6.1 Summary of the reciprocal best hit process for identifying possible viral orthologues of human IMPI genes encoding the mitochondrial proteome.

For genes with identified reciprocal best hits in at least one virus, the predicted protein domain structure of the viral predicted orthologues protein was compared to the structure of the human protein using NCBI CDD (Conserved Domain Database) searches (Marchler-Bauer *et al.* 2017). Proteins with functionally equivalent domains were kept as predicted orthologues.

Information on viruses

A list of viral strains with fully sequenced genomes was retrieved from the NCBI genome website (www.ncbi.nlm.nih.gov/genome), using the taxid 10239 for ‘Viruses’. Baltimore classification for each virus with a full genome was identified by retrieving the genome type from the NCBI taxonomy database (www.ncbi.nlm.nih.gov/taxonomy) and matching this to the Baltimore classification system (Baltimore 1971). The viral families for each viral strain were also retrieved, where available, from the NCBI taxonomy database.

The host phyla for each virus with at least one predicted orthologue of a human IMPI protein was identified from the NCBI taxonomy database, with additional hosts retrieved from the Virus-Host database where possible (Mihara *et al.* 2016).

Functional enrichment

Pathway annotation and enrichment was completed using the DAVID functional annotation tool (Huang *et al.* 2009a), using the IMPI gene list as a background, investigating annotation from KEGG, Reactome and GO Biological Process.

Phageness

‘Phageness’ (Kristensen *et al.* 2011) was calculated for each IMPI 2017 gene with at least one predicted viral orthologue, as follows:

$$\text{Phageness quotient} = \log \left(\frac{\text{Frequency of gene in sequenced viruses}}{\text{Frequency of gene in studied cellular organisms}} \right)$$

Where the frequency of genes in sequenced viruses is the fraction of viruses with fully sequenced genomes with a predicted orthologue of the gene; and the frequency of genes in studied cellular organisms is the fraction of the 359 species included in the orthologue database (*Chapter 2*) with a predicted orthologue of the gene.

Matrix clustering

R was used to cluster genes by their phylogenetic patterns. A binary matrix of the presence and absence of predicted orthologues of IMPI 2017 genes with at least one predicted viral orthologue was created. A distance matrix was calculated from this binary matrix using the function *dist()*, with the method ‘*euclidean*’. Clustering of the genes was achieved using the function *hclust()*, with the method ‘*ward.D2*’.

Localisation

Localisation of protein products in humans was taken from the COMPARTMENTS subcellular localisation database (Binder *et al.* 2014). If a compartment outside of the mitochondria had a score of 3 or over, the protein was assigned dual localisation.

Gene families

Gene families for each IMPI 2017 gene with a predicted viral orthologue were retrieved from the HGNC (HUGO Gene Nomenclature Committee) gene family dataset (Gray *et al.* 2016).

MTFP1 phylogenetic tree

Orthologues of *MTFP1* in cellular organisms were retrieved from the orthology dataset described in *Chapter 2*. Sequences of a cross-section of these species and all identified viral sequences were aligned using MUSCLE with default settings (Edgar 2004a, 2004b) and alignments were manually improved in Jalview (Waterhouse *et al.* 2009). A phylogenetic tree

was calculated using PhyML 3.0 using the best method (highest maximum likelihood score) from the two topology search methods (SPR – Subtree Pruning and Regrafting; and NNI – Nearest Neighbour Interchange) and 100 bootstrap replicates. The tree was visualised in Interactive Tree Of Life (Letunic & Bork 2016).

Viral mitochondrial carrier analysis

To create the tree of predicted viral mitochondrial carrier proteins, I aligned the viral protein sequences and the cow (*Bos taurus*) AAC1 sequence (accession: P02722.3) using MUSCLE with default settings (Edgar 2004a, 2004b) and manually assessed alignments in Jalview (Waterhouse *et al.* 2009). An unrooted phylogenetic tree was calculated using PhyML 3.0, using the best method (maximum likelihood score) from the two topology search methods (SPR – Subtree Pruning and Regrafting; and NNI – Nearest Neighbour Interchange) and 100 bootstrap replicates. The tree was visualised in Interactive Tree Of Life (Letunic & Bork 2016).

Matrix and cytoplasmic network motifs, as well as predicted contact points, were identified in each of the three repeats of the viral carriers from alignment with the cow (*Bos taurus*) AAC1 protein (accession: P02722.3) (Robinson *et al.* 2008). The consensus sequence for these motifs predicts the formation of six salt bridges from six pairs of residues – three forming a network on the side of the carrier closest to the cytoplasm and three forming a network on the mitochondrial matrix side of the carrier (*Figure 6.2*).

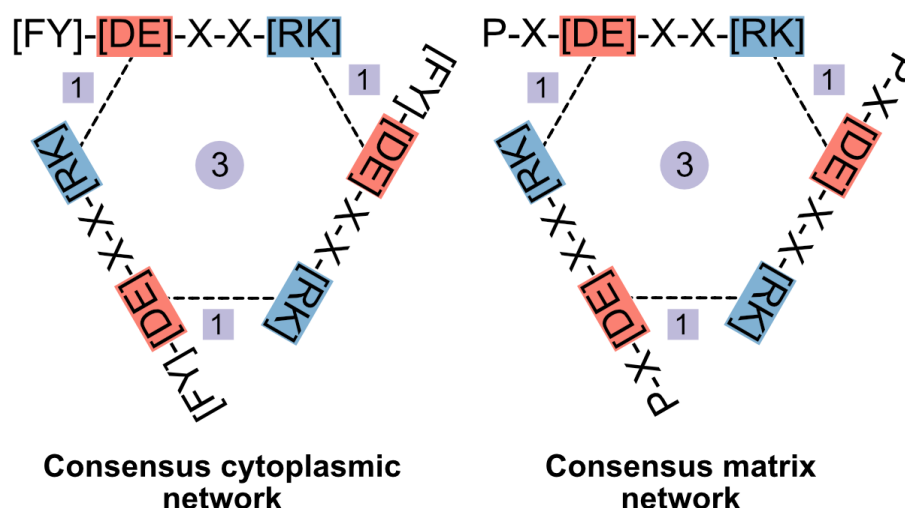


Figure 6.2 Scoring method for the strength of cytoplasmic and matrix salt bridge networks in mitochondrial carrier family proteins, using the consensus motifs for each network. Red = negative residue, blue = positive residue, dotted line = interaction. Numbers for each pair indicate the assigned strength of the interaction, whilst the central number indicates the total strength of the network.

Residues at these positions in the viral carriers were used to predict scores for the strength of the two networks, by summing the scores given for each of the three predicted pairs in each network.

Scores were assigned as follows (as in Robinson *et al.* 2008):

- A score of 1 was given for a residue pair predicted to form an ionic salt bridge,
- A score of 0.5 was given for a residue pair predicted to form a cation- π interaction or a hydrogen bond,
- A score of 0 was given for any other residue pairs.

Statistics

Statistical analyses were carried out in R. Statistical tests used, along with any correction for multiple testing, are indicated in the text.

Results

Identifying orthologues of human mitochondrial genes in viruses

Though there have been studies of individual genes shared between viruses and cellular organisms, there has not been a wide-scale look at genes shared between the human mitochondria and viruses. To investigate this, I first identified orthologues of human mitochondrial genes (as defined by both IMPI 2017 and MitoCarta 2.0) in viral sequences, using reciprocal best hit analysis. I used a liberal E-value of 10 as the cut-off for calling the orthologues, to minimise rates of false negatives for viral proteins that may be evolutionarily distant from the human sequence. This identified 8,263 viral orthologues of 649 IMPI 2017 genes across 3,559 viral strains, and 6,257 viral orthologues of 492 MitoCarta 2.0 genes across 3,047 viral strains. I then removed some predicted orthologues which did not show the same protein domain structure as the human protein. E-values of retained and removed hits overlapped substantially (*Figure 6.3*), with some retained hits having E-values as large as 2.5.

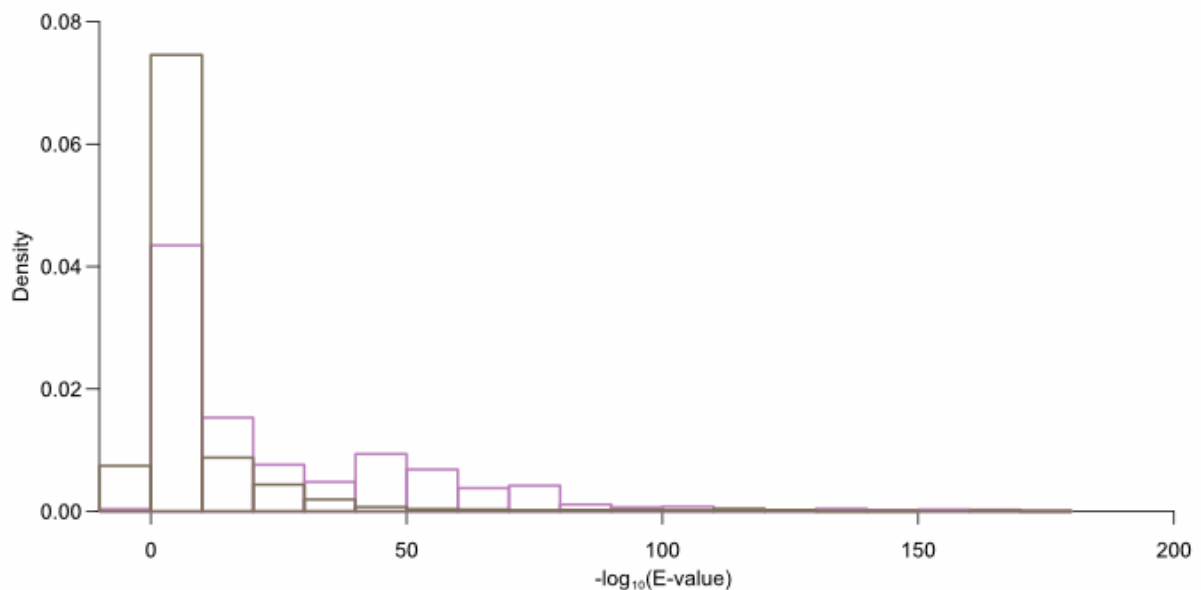


Figure 6.3 Comparative histograms of \log_{10} (reciprocal BLASTp E-values) for retained (purple) and removed (grey) predicted viral orthologues of human IMPI mitochondrial proteins.

After domain assessment, there were 5,204 predicted viral orthologues of 156 IMPI 2017 genes (10.1% of all genes), across 2,540 viral strains, and 4,274 predicted viral orthologues of 125 MitoCarta 2.0 genes (10.8% of all genes), across 2,263 viral strains.

There is a large overlap of genes between IMPI 2017 and MitoCarta 2.0, and this is reflected in the strong correlation between the number of orthologues found per virus in each data set, only including those with full genomes recorded by the NCBI (*Figure 6.4*, correlation coefficient = 0.936, $p < 2.2 \times 10^{-16}$). The number of orthologues predicted per virus tends to be slightly higher using the IMPI 2017 dataset, probably reflective of the larger size of the gene set compared to MitoCarta 2.0. Therefore, from this point on, all analyses use the IMPI 2017 dataset to define the genes of the human mitochondrial proteome.

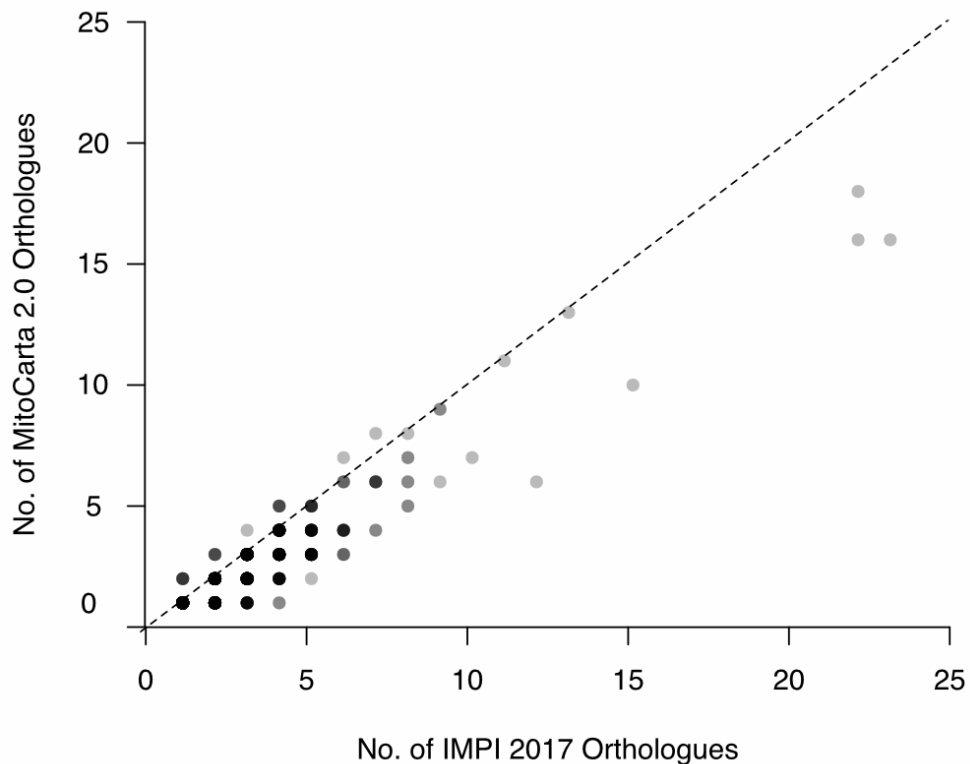


Figure 6.4 Comparison of the number of predicted mitochondrial gene orthologues per viral strain, between those defined by IMPI 2017 and those defined by MitoCarta 2.0. Darker shades represent a larger number of viruses.

Viruses with predicted mitochondrial gene orthologues

I then investigated the composition of this dataset of viral orthologues of human mitochondrial genes. To investigate the type of viruses most likely to have a predicted orthologue of a mitochondrial protein, I classified all viral strains with full genomes in NCBI using the Baltimore classification system, which separates viruses by the type of genome and method of genome replication (Baltimore 1971). I then identified the number of viral strains within each group which had at least one predicted mitochondrial orthologue (*Table 6.1*).

Table 6.1 Predicted mitochondrial protein orthologues in viral strains with fully sequenced genomes, separated by Baltimore classification (*ds* = double-stranded, *ss* = single-stranded, *RT* = reverse-transcribing).

Baltimore classification		No. of viruses with mitochondrial orthologue(s)	Total no. of viruses	% of viruses with mitochondrial orthologue(s)
Group	Genome			
I	dsDNA	1285	2736	47.0
II	ssDNA	0	1003	0.0
III	dsRNA	0	300	0.0
IV	(+)ssRNA	37	1313	2.8
V	(-)ssRNA	0	442	0.0
VI	ssRNA-RT	4	66	6.1
VII	dsDNA-RT	6	86	7.0

There are predicted viral orthologues of IMPI 2017 genes in viruses with both DNA and RNA genomes, as well as both double-stranded (ds) and single-stranded (ss) genomes. Viral strains with dsDNA genomes have by far the highest number of strains with at least one predicted mitochondrial orthologue, as well as the highest percentage of all sequenced strains with at least one predicted mitochondrial orthologue – nearly half of all sequenced dsDNA viruses. Few members of the other Baltimore groups have predicted orthologues of mitochondrial genes, with less than 10% of the total fully sequenced viral strains having at least one predicted orthologue.

The viral strains with predicted orthologues of mitochondrial genes infect a wide range of host species, including bacteria, archaea and eukaryotes (*Figure 6.5*). This is perhaps reflective of the wide spread of IMPI 2017 orthologues across all types of cellular organisms (*Chapter 2*). Viral host phyla which include popular model organisms (e.g. *Escherichia coli* in Proteobacteria) or other well-studied species (such as humans) have higher numbers of

viruses with predicted mitochondrial orthologues. This suggests that the viral proteome analysed is skewed towards certain types of viruses. Very few of the viral strains with an identified mitochondrial orthologue infect archaeal species.

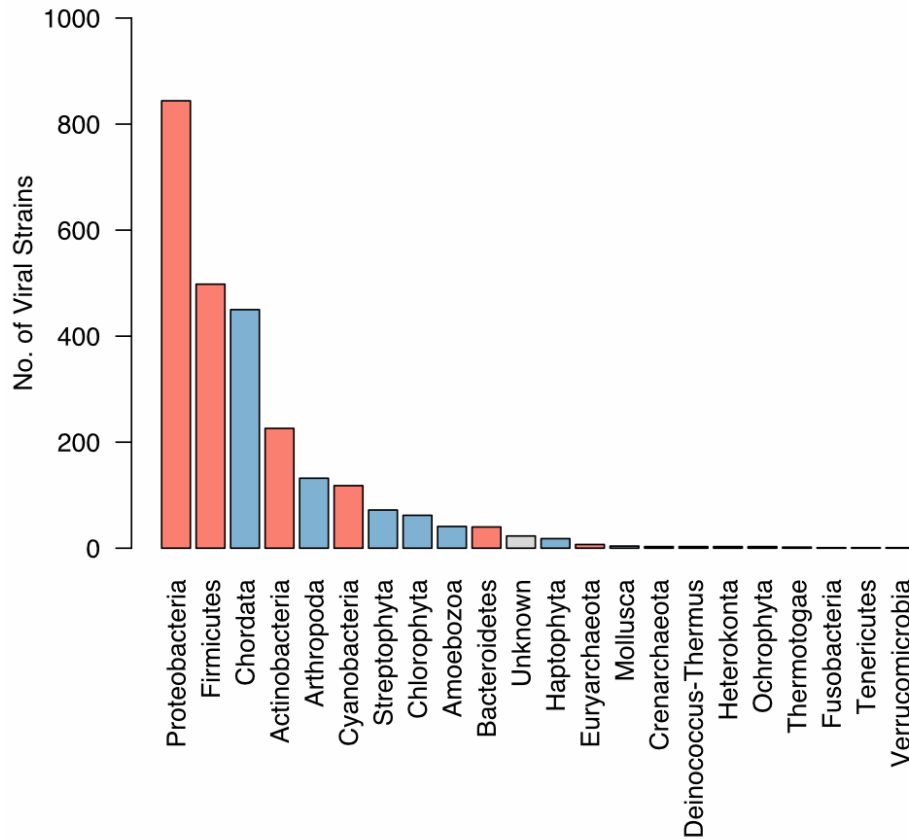


Figure 6.5 Host phyla (or closest available taxonomic rank) for viral strains with at least one predicted orthologue of IMPI mitochondrial proteins. Colours indicate prokaryotic (red), eukaryotic (blue) or unknown (grey) hosts.

Viruses with a fully sequenced genome and at least one predicted orthologue of an IMPI 2017 gene belong to twenty-two different viral families (Figure 6.6), with an additional 74 of these viruses unclassified. This is around 20% of all viral families with at least one fully sequenced member strain. The percentage of viral strains within each family with a predicted mitochondrial orthologue ranges from 6.1% (Retroviridae) to 100% (several different families, with between 1 and 24 fully sequenced viral members). The largest number of mitochondrial protein orthologues are predicted in ‘*Acanthamoeba polyphaga mimivirus*’ (NCBI taxid: 212035), a member of the Mimiviridae – one of several families of so-called ‘giant viruses’.

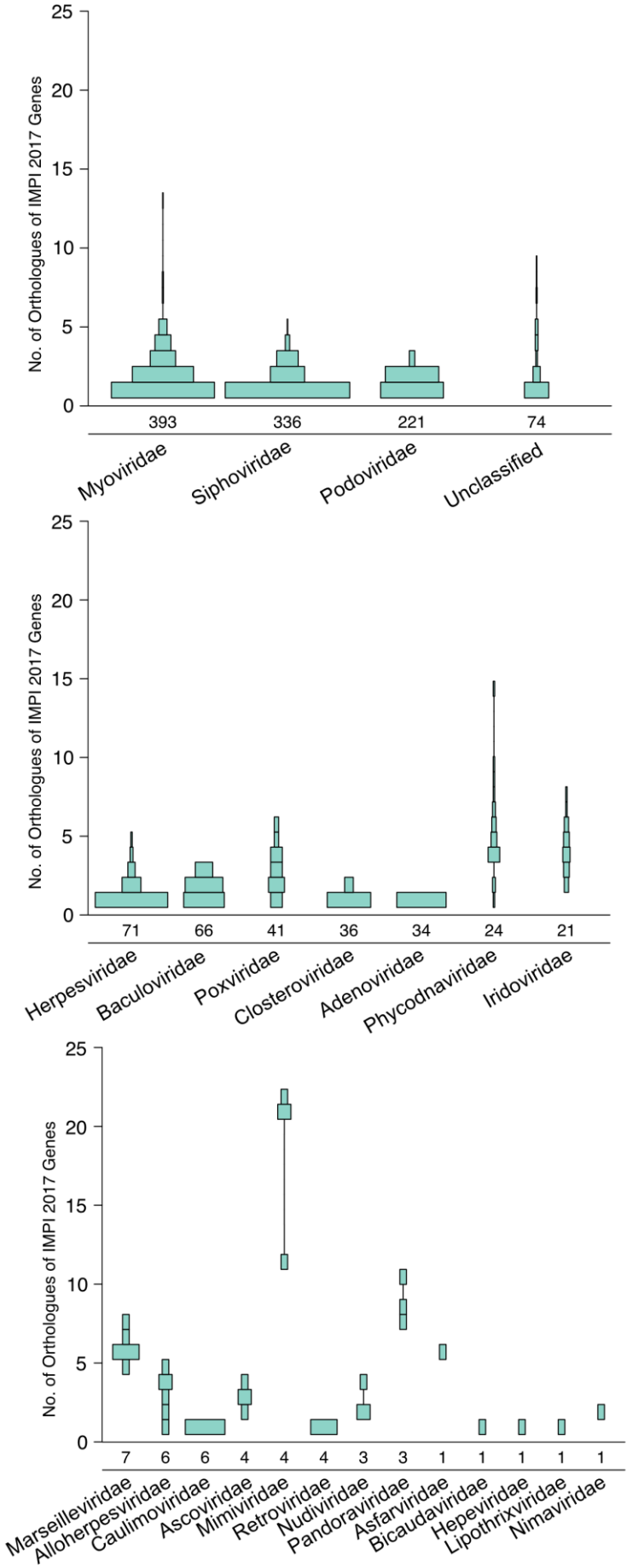


Figure 6.6 (overleaf) *Number of mitochondrial orthologues in viruses with full genomes separated by viral family. Numbers under each plot indicate the total number of viruses in that family with at least one mitochondrial orthologue. Width of each bar indicates the number of viruses with the indicated number of orthologues (note the different scales between the plots, to account for the variation between the number of viral strains in different viral families).*

In summary, a wide range of viruses contain orthologues of human mitochondrial genes. They are not limited by genome type, host cell or the viral family, though there are examples of each which currently have no predicted viral orthologues of mitochondrial genes.

Function of viral orthologues of mitochondrial proteins

I then explored the types of mitochondrial proteins which viruses share with humans. 256 different IMPI 2017 genes have a predicted viral orthologue (*Appendix III – Table 1*). After grouping the genes into related HGNC protein families, there were 61 different protein families identified, as well as an additional 62 genes with no associated family. Thus, the viral proteome includes orthologues of the products of genes producing a wide-ranging set of mitochondrial proteins.

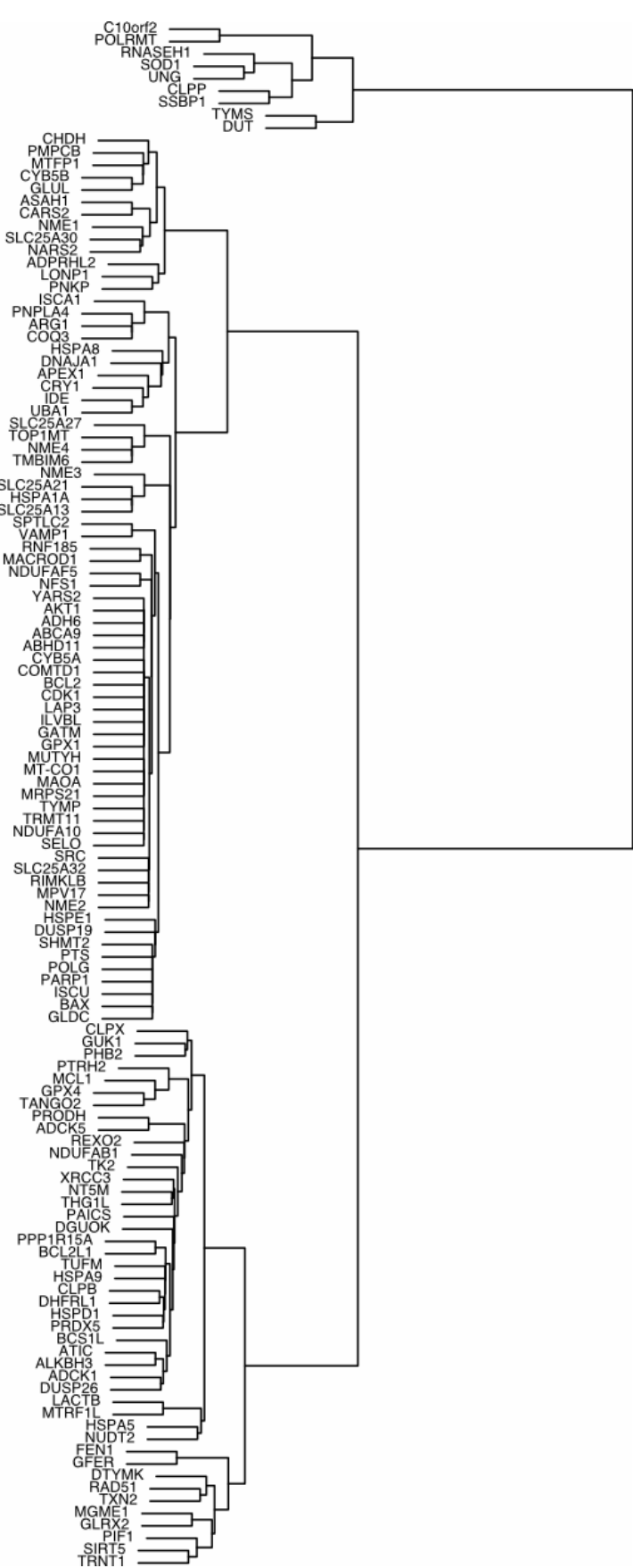
I assessed the enrichment of functional terms in the mitochondrial genes with viral orthologues, compared to IMPI 2017. Four terms were enriched with $p < 0.05$ after correction for multiple testing (*Table 6.2*). These categories are important for correct transcription and/or translation in humans, which suggests viruses share genes with humans encoding mitochondrial proteins to assist with their replication, perhaps through hijacking of the mitochondrial transcription/translation system.

Table 6.2 Functional annotation enrichment for mitochondrial genes with at least one viral orthologue. Terms from KEGG, GO Biological Process (GO BP) and Reactome with $p < 0.05$ after multiple testing correction.

Type	Term	Count	%	p -value	Benjamini corrected p -value
KEGG	Pyrimidine metabolism	12	7.7	8.8×10^{-7}	1.5×10^{-4}
GO BP	tRNA aminoacylation for protein translation	10	6.5	4.5×10^{-5}	2.7×10^{-2}
GO BP	DNA repair	11	7.1	2.6×10^{-5}	3.1×10^{-2}
GO BP	DNA replication	8	5.2	1.2×10^{-4}	4.8×10^{-2}

Clustering genes by their pattern of orthologues from viruses with full genomes, the nine most frequently identified genes cluster together (*Figure 6.7*).

Figure 6.7 Dendrogram of the IMP1 2017 genes with at least one predicted viral orthologue, clustered by orthologue prediction pattern in fully sequenced viruses.



These nine genes are each identified in over 150 viral strains (*Table 6.3*). However, of the 2,112 viral strains with a predicted orthologue of at least one of these genes, no virus was identified that encodes more than five of these genes, and 65.4% of viral strains only encode one of these genes. This suggests that the clustering of these genes is due mainly to the larger number of viral strains with identified orthologues, rather than a true relationship between the presence of any of these genes in a viral genome.

Table 6.3 *IMPI genes with predicted orthologues in over 150 viral strains.*

Ensembl ID	Gene	Localisation	Function
ENSG00000167136	<i>DUT</i>	Dual	Production of dUMP in thymidine nucleotide anabolism and maintenance of low dUTP/dTTP ratio
ENSG00000176890	<i>TYMS</i>	Dual	Thymidine nucleotide anabolism
ENSG00000106028	<i>SSBP1</i>	Dual	Binds single-stranded DNA during replication
ENSG00000125656	<i>CLPP</i>	Mitochondria	ATP-Mg ²⁺ dependent peptidase
ENSG00000099821	<i>POLRMT</i>	Mitochondria	Mitochondrial transcription
ENSG00000171865	<i>RNASEH1</i>	Dual	Degradation of RNA-DNA hybrids during transcription
ENSG00000076248	<i>UNG</i>	Dual	Removal of uracil from DNA
ENSG00000142168	<i>SOD1</i>	Dual	Free radical metabolism
ENSG00000107815	<i>C10orf2</i>	Mitochondria	DNA helicase involved in replication

Three of the top identified genes are involved in thymidine nucleotide metabolism, particularly related to uracil (*Table 6.3*). Controlling uracil metabolism may be important to maintaining the integrity of the viral genome, as DNA polymerases cannot necessarily distinguish between thymine and uracil during replication (Brynolf *et al.* 1978; Tye *et al.* 1978). However, *DUT* has also been shown to play additional roles in the lifecycle of several viruses. For example, the Epstein-Barr virus orthologue of *DUT* helps control the immune response by modulating host-cell inflammatory cytokines (Ariza *et al.* 2009).

The proteins encoded by most of these top predicted genes are thought to be dual localised in humans, suggesting a potential weaker link to controlling the function of the mitochondria. Two of the three genes with proteins mainly localised to the mitochondria are involved in DNA replication and transcription of DNA producing RNA. These two genes – mitochondrial RNA polymerase (*POLRMT*) and mitochondrial DNA helicase (*C10orf2*) – are thought to be of phage origin (Masters *et al.* 1987; Cermakian *et al.* 1996; Spelbrink *et al.* 2001).

Overall, the viral proteome contains potential orthologues of a wide range of mitochondrial genes. The functions of these genes, including those which are most frequently predicted as orthologues, tend to be linked to nucleotide metabolism, transcription and translation.

Spread of viral orthologues of mitochondrial proteins

Another way to look at these results is to look at the spread of orthologues in the studied viruses. The phageness quotient of a gene is a log-odds ratio measure of the frequency of a gene in viruses compared to the frequency of a gene in another set of organisms (Kristensen *et al.* 2011) – in this case, the cellular organisms used to create an orthologue database in *Chapter 2*. A gene which is more frequently identified in viruses than cellular organisms would have a score above zero, whilst a gene which is more frequently identified in cellular organisms than viruses would have a score below zero. For the 994 IMPI 2017 (86.4%) genes with no predicted viral orthologues, the phageness quotient would be $-\infty$. All the genes with at least one viral orthologue are more frequently identified in the studied cellular species than the sequenced viruses – i.e. they all have scores below zero (*Figure 6.8*).

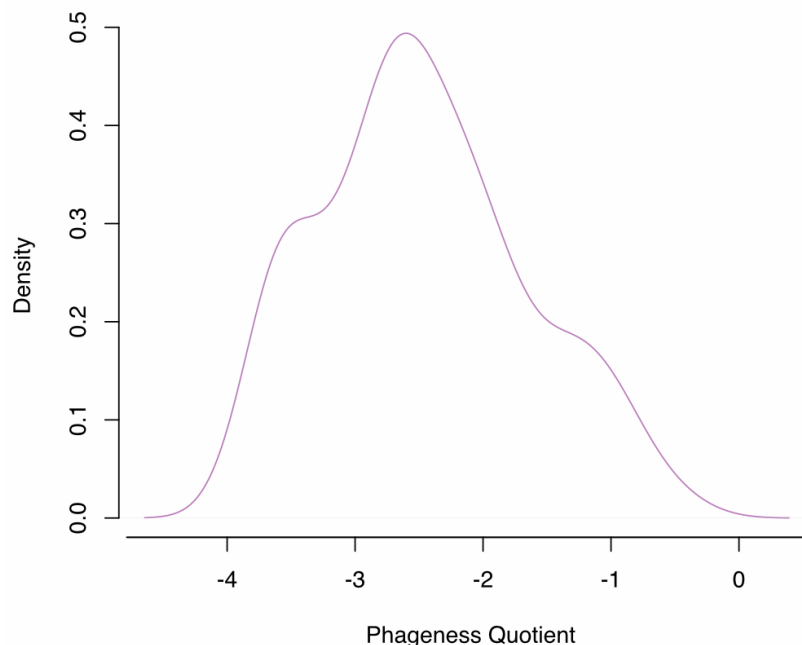


Figure 6.8 Density plot of the phageness quotient of each of the IMPI 2017 genes with at least one predicted viral orthologue. Gene phageness quotient is a log-odds ratio measure of the frequency of gene orthologue detection in all fully sequenced viruses compared to the frequency of gene orthologue detection in the cellular organisms included in the orthology dataset from Chapter 2.

There is a peak at a phageness quotient of approximately -2.5, which is equivalent to a gene that is approximately 300 times more frequent in the studied cellular organisms than viruses. This suggests that though there a wide variety of mitochondrial proteins with predicted orthologues in viruses, each gene is relatively rare compared to their spread in cellular organisms.

One explanation for this could be that viral orthologues of mitochondrial genes are not spread widely across families of viruses but instead tend to be unique to certain viral families. This is supported by looking at the number of viral families with an identified orthologue of each protein, with most proteins having predicted orthologues in very few families (*Figure 6.9*). The two most widespread IMPI 2017 genes across viral families are *DUT* (deoxyuridine triphosphatase, identified in 19 families) and *TYMS* (thymidylate synthetase, identified in 13 viral families). However, over half (56.2%) of all the genes with at least one predicted viral orthologue are only identified in one viral family. A possible explanation for this would be the unique transfer of genes between a host and a virus, with spread throughout a family as the viral family expands, although this idea would need to be supported by further data.

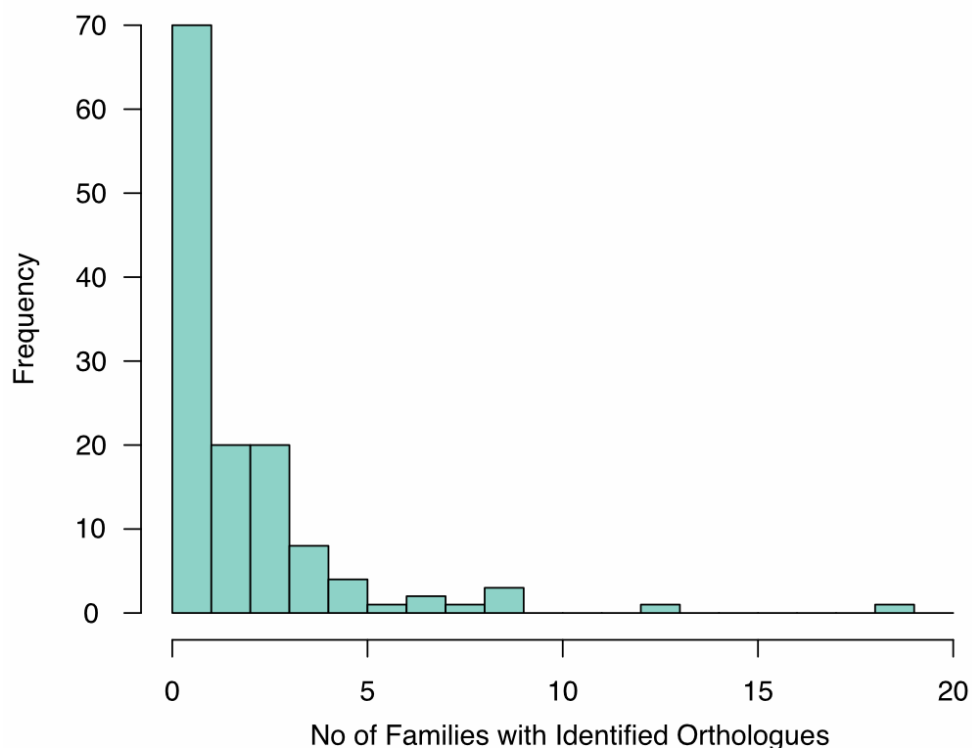


Figure 6.9 Histogram of the number of viral families with a predicted orthologue, for each IMPI 2017 gene with at least one viral orthologue and one virus with an assigned family.

Transfer of genes between virus and host genomes

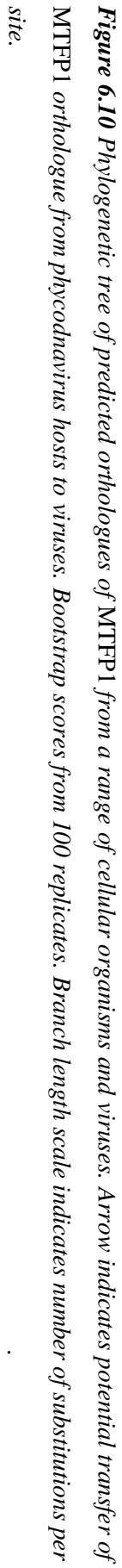
Proteins which are shared between viruses and the hosts may indicate gene transfer between these two entities. There is already evidence for this type of transfer with, for example, the mitochondrial RNA polymerase (*POLRMT*) (Cermakian *et al.* 1996). I decided to use phylogenetic trees to investigate the possible origin of some mitochondrial genes encoded by viruses. Clustering of viral sequences close to their hosts, rather than separately from cellular organisms, would provide evidence that there was possible transfer of the gene from the host to the virus or vice versa.

One gene I investigated was *MTFPI* (Mitochondrial Fission Process 1), which encodes a protein regulating the fission of mitochondria (Tondera *et al.* 2004, 2005). This gene was a good candidate for analysis, as it was only identified in eukaryotic species in the IMPI 2017 orthologue database (*Chapter 2*) and contains a functional domain (pfam10558) unique to this gene in humans. This could reduce confusion due to potential gene transfer between prokaryotic species and/or presence of paralogues. Viral orthologues of these gene were predicted for seventeen viral strains, from two different families – the Mimiviridae (fourteen sequences) and the Phycodnaviridae (three sequences).

I used PhyML to calculate a phylogenetic tree of a selection of eukaryotic and all viral orthologues of human *MTFPI* (*Figure 6.10*). The tree suggests that there may be two separate origins of this gene in viruses. The Phycodnaviridae sequences cluster with the green algae, particularly *Ostreococcus* and *Micromonas* species – these are the hosts of the detected viruses encoding these sequences. This provides some evidence that the genes in the Phycodnaviridae may originate from horizontal transfer from their host cells. (Whilst it is possible that the gene moved the opposite way – from virus to host cell – it is unlikely considering the spread of the gene throughout the eukaryotes).

However, the sequences from the Mimiviridae cluster separately and seem distantly related to all other sequences. This phylogenetic tree, therefore, does not provide evidence about a possible origin of these genes in the Mimiviridae. There are several possibilities for the lack of strong relationship between these sequences and the rest of the phylogenetic tree, including that Mimiviridae sequences have become highly diverged from the original sequence or that there are no sequences included from species close to the Mimiviridae true hosts, as these are not always clear (Claverie *et al.* 2009).

Though I also investigated several other genes (including *BCSIL*, *THGIL*, *PRODH*, and *GFER*) no other phylogenetic tree produced such a clear result of possible host/viral gene transfer as that seen for *MTFPI*. This could be due to some of the limitations I tried to avoid when choosing *MTFPI* for analysis, such as having prokaryotic orthologues or being part of a larger gene family, which are harder to distinguish from each other. It could also be a limitation of the technique, which is dependent on the sequences included and the rate of divergence of the sequences.



Mitochondrial carriers in viruses

The mitochondrial carrier family is a group of proteins with a similar tri-repeat structure (Gutiérrez-Aguilar & Baines 2013). Most of the characterised members of this family move substrates across the inner mitochondrial membrane, though there are exceptions to this rule. A single viral orthologue of a mitochondrial carrier from a member of the Mimivirus group has been expressed and characterised as a dATP and dTTP carrier (Monné *et al.* 2007). I used my viral mitochondrial protein orthologue dataset to identify additional viral strains with predicted orthologues of mitochondrial carriers.

There are 27 viral mitochondrial carrier orthologues in the orthology dataset across 26 viral strains. One viral strain (*Hokovirus HKV1*) encoded two predicted mitochondrial carrier proteins. These viral strains belonged to the families Mimiviridae, Nudiviridae and Hytrosaviridae, as well as some unclassified viruses. Mimiviridae have been proposed as a member of a new order of viruses known as the Megavirales (Colson *et al.* 2013), whereas the other families have not been classified into a viral order. They are all double-stranded DNA viruses, with some of the largest known viral genomes (<http://www.giantvirus.org/top.html>). The host species of the Nudiviridae and Hytrosaviridae are insects and crustaceans, whereas most Mimiviruses have been cultivated in amoeboid species, but their true hosts are not characterised (Claverie *et al.* 2009).

I used the sequences of these predicted viral carriers to produce an unrooted phylogenetic tree (*Figure 6.11*).

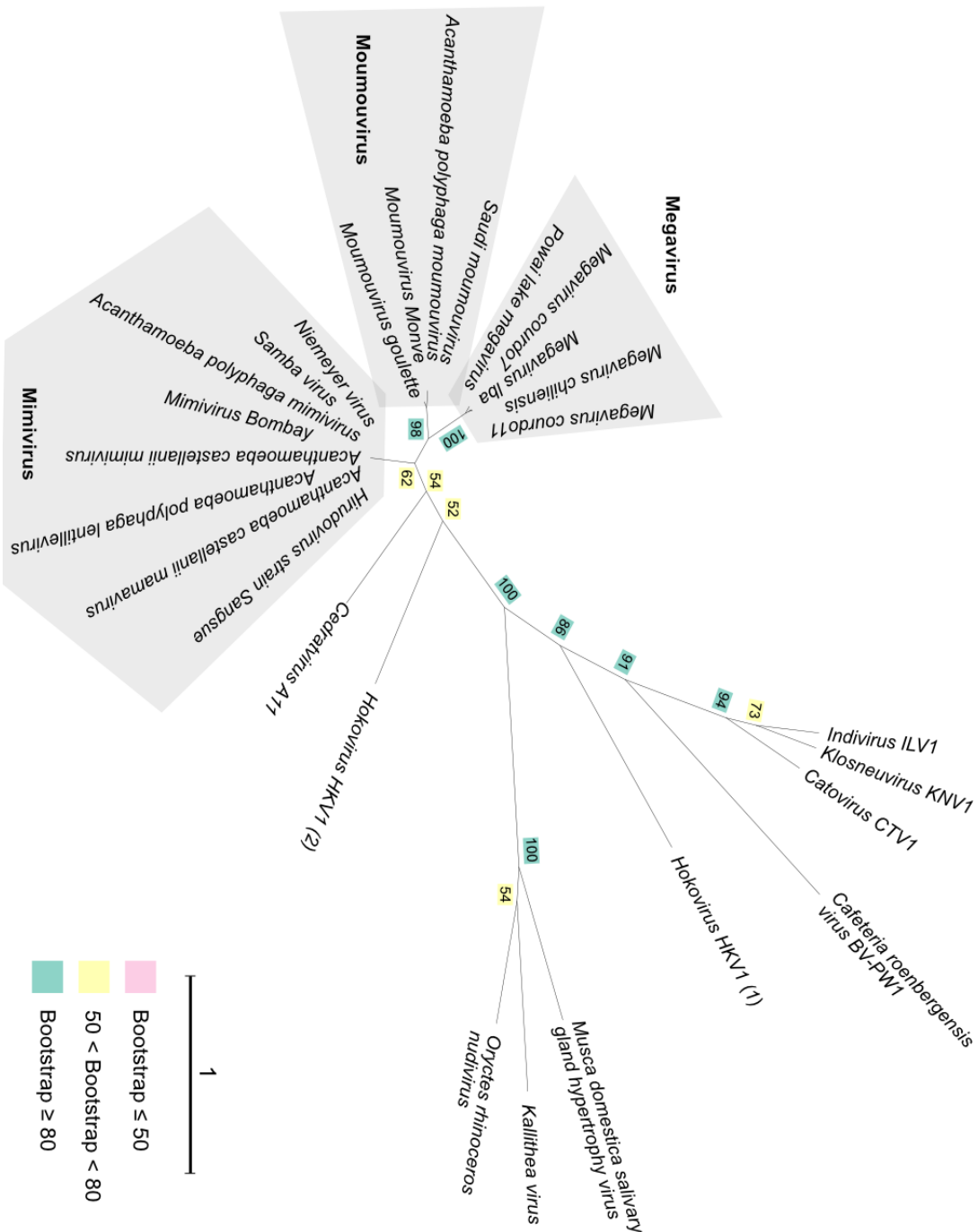


Figure 6.11 Tree of predicted viral orthologues of mitochondrial carrier family proteins. Bootstrap scores from 100 replicates. Branch length scale indicates number of substitutions per site.

The sequences from the three members of the Nudiviridae and Hytrosaviridae cluster together (bootstrap 100/100), with the remaining sequences grouping into two additional clusters, each one containing one of the sequences from *Hokovirus HKV-1*. It is therefore possible that these mitochondrial carriers are functioning differently from the carriers encoded by Megavirales members.

This is possibly supported by looking at the sequences of the viruses from the different families. The three viruses from the Nudiviridae and Hytrosaviridae encode different residues at the substrate contact points II and III – which define the likely substrate specificity of the carriers – compared with the other viral mitochondrial carriers (*Figure 6.12*). This may suggest that they transport a different substrate from other viral mitochondrial carriers (dATP for the characterised transporter from Mimivirus (Monné *et al.* 2007)), or that the substrate binds differently.

Mitochondrial carriers function by the formation and breakage of salt bridge networks – one at the mitochondrial matrix face of the protein, and the second at the intermembrane space or cytoplasmic face of the protein (King *et al.* 2016). With the increased number of viral mitochondrial carrier sequences available, it becomes possible to assess the average predicted strength of the mitochondrial and cytoplasmic salt networks (Robinson *et al.* 2008). On average, the mitochondrial carriers from the Nudiviridae and Hytrosaviridae have both a reasonably strong mitochondrial (1.8/3) and cytoplasmic (1.7/3) salt bridge network, suggesting these transporters function as exchangers. The other viral mitochondrial carriers have a strong mitochondrial salt bridge network (2.8/3) and a weaker cytoplasmic salt bridge network (1.2/3) (*Figure 6.12*). Though this suggests these carriers may be uniporters, characterisation of the Mimivirus carrier showed only exchange activity (Monné *et al.* 2007).

In summary, this analysis predicts there may be at least two groups of viral mitochondrial carriers in different viral families, with potentially different substrate binding and exchange activities.

	Repeat 1				Repeat 2				Repeat 3				Matrix Network Score	Cytoplasmic Network Score
	Matrix Signature Motif	Contact Point I	Cytoplasmic Signature Motif		Matrix Signature Motif	Contact Point II	Cytoplasmic Signature Motif		Matrix Signature Motif	Contact Point III	Cytoplasmic Signature Motif			
Mimivirus	P I C T F K	Q	F R F F E		P I D N I K	K	E T L N		P L D F L K	R	I D F L N		2.5	0.5
Moumouvirus	P I C T F K	Q	Y R Y R R		P I D A I K	K	Y D K I N		P L D Y L K	R	I D F L N		2.5	1.5
Megavirus	P I C T I K	Q	Y R W F E		P I D T I K	K	F D L F E		P L D Y L K	R	I D Y L N		2.5	0
Cedratvirus A11	P I C T I K	Q	Y R R L Q		P L D V I K	K	F D L F E		P I D Y M K	R	I K F I E		2.5	2.5
Indivirus ILV1	P F D T I K	V	Y T N T Y		P V E R I K	R	Y E G F K		P Q D C I K	R	F E Y L K		3	1
Kiosneuvirus KNV1	P F D T I K	V	Y K T L D		P V E K I K	R	Y E N M K		P Q D C I K	R	M D L L K		3	1
Catovirus CTV1	P F D T I K	V	F S M L N		P Y E R I K	R	Y E Y L K		P Q D R I K	R	M E L L K		3	1.5
Hokovirus HKV1 (1)	P F D T L K	Q	K Y L L		P F D V V K	R	D L L R		P L D V I K	R	V T L F N		3	0.5
Hokovirus HKV1 (2)	P I C T T K	Q	Y E Y V K		P F D Y A K	R	N D Q F N		P L D F I R	R	V E L I K		2.5	2.5
Cateteria roenbergensis virus BV-PW1	P L D S L K	N	N L M E		P I E Y F K	R	R N F K		P L D T I K	R	Y V K L T		3	1
Average													2.8	1.2
Nudiviridae/Hytrosaviridae														
Musca domestica salivary gland hypertrophy virus	P L D V L R	A	L N D Y L		P L W T I K	N	Y D I F K		P I F A M R	Q	F E A I K		2	1
Oryctes rhinoceros nudivirus	P F D T I R	Y	N R H F K		P L W T I K	N	Y D I C K		P I F A L R	Q	Y E K F K		2	2
Kallithea virus	P L D T I R	F	N R Y L K		P L W V M K	L	Y D I L K		P L L A L R	Q	Y E N S V		1.5	2
Average													1.8	1.7

Figure 6.12 Key features from the three mitochondrial carrier repeats of the viral mitochondrial carriers. Colours represent pairs of positions which are predicted to form salt bridges in fully formed networks. Network scores for the matrix and cytoplasmic networks are summed from the three pairs in each network, given: a score of 1 for a positive/negative pair; a score of 0.5 for pairs predicted to form cation- π or hydrogen bond interactions; and a score of 0 for any other interaction. Predicted substrate contact points are also pictured. A consensus sequence is shown for the viruses from the three viral

Discussion

The scientific literature is full of examples of viral proteins which have been shown to influence the function of a host cell mitochondria and/or seem to be orthologous with mitochondrial protein sequences (Ohta & Nishiyama 2011). However, I could not find evidence of a wide-scale screen for potential mitochondrial protein orthologues in viruses, which could help our understanding of the influence of mitochondria in viral infections. Therefore, I used reciprocal best hit analysis to identify possible orthologues of human mitochondrial proteins across the sequenced viral proteome.

This analysis showed that viruses share a wide range of human mitochondrial proteins, with around 10% of IMPI 2017 genes having at least one predicted orthologue. These genes are enriched for functions which can be linked to the success of viral replication (*Table 6.2*), including nucleotide metabolism, DNA replication/repair, and translation. Viruses have been increasingly linked to mitochondrial metabolism in general, including key aspects such as the tricarboxylic acid cycle (Claus & Liebert 2014). There has also been a described case of an entire metabolic pathway transferring from an algal host to a virus (Monier *et al.* 2009). Despite this, very few human mitochondrial metabolic enzymes, outside of those involved in nucleotide metabolism, had predicted orthologues within the studied viral proteome.

The predicted orthologues are widely spread throughout the viruses, with approximately 22% of fully sequenced viruses encoding orthologues of at least one gene linked to the mitochondria in humans. Viruses with double-stranded DNA genomes were the most likely to encode an orthologue of a mitochondrial protein (~ 45% of sequenced viruses). Orthologues of individual genes are often unique to certain families of viruses (*Figure 6.9*), which perhaps supports the idea of individual cases of gene transfer between different types of viruses and host cells.

The viruses with the largest numbers of predicted mitochondrial orthologues were the so-called ‘giant viruses’ – those known to have much larger than average genomes – with, for example, 32 orthologues predicted in the unclassified *Klosneuvirus KNV1*. This would perhaps be expected, as the traditional idea is that giant viruses grew due to the acquisition of a variety of genes from their hosts (Koonin 2005b). However, more recent work has predicted that this type of genetic transfer is not as common as had been expected (Monier *et al.* 2007). It is therefore interesting that so many mitochondrial proteins seem to be shared between giant

viruses and cellular organisms. Mitochondria may be key to the replication of some giant viruses, with mitochondria surrounding the cytoplasmic virus factories necessary for replication (Suzan-Monti *et al.* 2007; Campos *et al.* 2014). It may be that these orthologues of mitochondrial proteins help recruit or co-opt some function of these mitochondria.

Genes shared between viruses and cellular organisms are often assumed to be the result of horizontal gene transfer – either from virus to host cell or vice versa. This phenomenon has been observed in a wide variety of viruses and genes (Liu *et al.* 2010, 2011; Gilbert *et al.* 2016). *MTFPI* orthologues in members of the Phycodnaviridae (more specifically, Prasinoviruses) are a good example of possible horizontal gene transfer from the host cell (in this case, marine green alga) to the virus. Prasinoviruses are not known to specifically utilise the mitochondria during their lifecycle (Weynberg *et al.* 2017). *MTFPI*, however, has been linked to cell survival and apoptosis in eukaryotic cell lines (Morita *et al.* 2017) and during the treatment of cancer (Aung *et al.* 2017a, 2017b). It would therefore be interesting to explore the expression of the *MTFPI* orthologues from the Prasinoviruses, particularly in relation to how they may control apoptosis of the host cell during the viral life cycle.

Exploration of the mitochondrial carrier orthologues encoded by viruses has identified at least three different clusters of related sequences. One cluster contains the mitochondrial carrier encoded by the *Mimivirus*, which has previously been shown to transport dATP. Another cluster contains sequences from the Hytrosaviridae and Nudiviridae – viruses which infect insects. Sequences of carriers from these clusters differ, both in the strength of their cytoplasmic/matrix salt networks and in their substrate binding sites. The mitochondrial carrier in *Musca domestica* salivary gland hypertrophy virus (a member of the Hytrosaviridae) is one of the top fifteen most-expressed transcripts during viral infection (Kariithi *et al.* 2017). Understanding the function and activity of this carrier in the Hytrosaviridae and Nudiviridae may improve understanding of the viral infective processes.

The variety of mitochondrial proteins with possible orthologues detected across the viral proteome suggests that viruses have developed myriad ways to control and co-opt the mitochondria and its processes, beyond those which have currently been studied.

Investigating the expression and importance of these viral genes could contribute to the understanding of the importance of mitochondria in viral infections and may illuminate new methods of controlling these processes.

Conclusions

The currently sequenced viral proteome contains predicted orthologues of a wide range of mitochondrial proteins. Viruses from a variety of different families and with different types of genomes encode predicted mitochondrial protein orthologues. Those genes with predicted viral orthologues are often related to nucleotide metabolism, DNA replication, transcription and translation – all functions which can be linked to the success of viral replication. Most genes are restricted to only one or a few viral families. There is some evidence for potential host cell to virus gene transfer for one of these genes (*MTFPI*); as well as for at least one more group of mitochondrial carrier orthologues, beyond those which have been functionally characterised, which could be of interest to study.

Chapter 7

Conservation of non-enzymatic

$S \Rightarrow N$ lysine acetylation

Introduction

Lysine N-acetylation

Proteins can be adorned with a variety of modifications. One type of protein modification is lysine *N*-acetylation (addition of an acetyl group). Acetyl groups can be enzymatically transferred from acetyl-CoA to the lysine ϵ -amino group by acetyltransferases and removed by deacetylases (Choudhary *et al.* 2014).

Sirtuin 3 (*SIRT3*) is an NAD^+ -dependent, mitochondrial deacetylase (Lombard *et al.* 2007), but there is no definitive mitochondrial acetyltransferase. The mitochondrial matrix is a unique environment, with high levels of acetyl-CoA (Hansford & Johnson 1975) and a high pH, due to H^+ ion movement out of the matrix which is necessary to generate the proton motive force (Llopis *et al.* 1998; Abad *et al.* 2004). Non-enzymatic acetylation of proteins has been demonstrated in these conditions (Wagner & Payne 2013).

Recently, work from the Murphy group showed that, at least *in vitro*, lysine residues can be non-enzymatically *N*-acetylated by acetyl-CoA. This occurs via a reversible process involving, first, nucleophilic attack of a cysteine residue on the acetyl-CoA forming an *S*-acetylated cysteine (James *et al.* 2017). In some cases, this acetyl group is transferred to a nearby lysine residue ($S \Rightarrow N$ acetyl-transfer; *Figure 7.1*), forming an acetylated lysine.

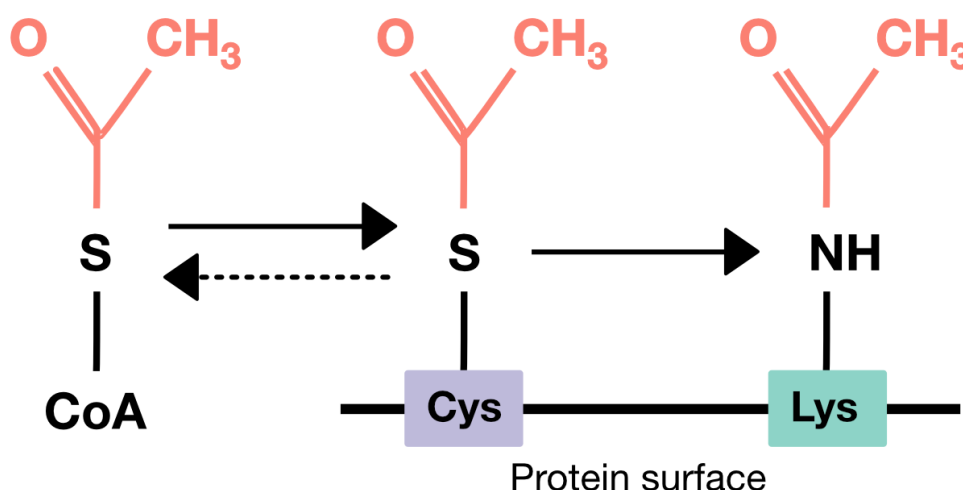


Figure 7.1 Non-enzymatic $S \Rightarrow N$ acetylation of surface lysine residues. Acetyl groups (red) are transferred from acetyl-CoA; first to an exposed cysteine residue on the protein surface, then onto a lysine residue which is structurally close.

Studies have estimated that ~35% of mitochondrial proteins have potential acetylation sites (Anderson & Hirschey 2012), but few have characterised changes in function due to acetylation (Baeza *et al.* 2016). Most of the studied changes are in key metabolic enzymes, where the acetylation causes inhibition and levels can be controlled by the NAD⁺-dependent *SIRT3* deacetylase (Baeza *et al.* 2016). This organised acetylation, therefore, may link metabolism to indicators of energy levels within the cell (acetyl-CoA and NAD⁺).

However, when lysine acetylation in the mitochondria is quantified by mass spectrometry, most acetylated residues have very low stoichiometry (< 1%) (Weinert *et al.* 2014, 2015). This brings up the question of how functional a sizeable proportion of the detected acetylation is.

One proposal is that some non-enzymatic acetylation acts as a ‘carbon stress’ on the cell, with higher concentrations of acetyl-CoA increasing non-specific acetylation (Wagner & Hirschey 2014). Acetylation may disrupt protein function through altering interactions with other proteins, substrates or cofactors, and/or increasing protein aggregation, which over time may damage the cell (Kuczyńska-Wiśnik *et al.* 2016). This could contribute to the link between acetylation and a variety of disease states, including heart failure (Lee & Tian 2015) and neurodegenerative disorders, such as Huntingdon’s disease, Alzheimer’s disease and Parkinson’s disease (Min *et al.* 2015).

Animal longevity

Animal lifespan correlates with size – larger animals live, on average, longer than smaller animals (Blueweiss *et al.* 1978). However, there is variation between different groups of animals and extraordinary exceptions within groups. For example, bats live longer, on average, than rodents and other non-flying eutherians of a similar size (Austad & Fischer 1991). However, *Heterocephalus glaber* (the naked mole-rat) is an exception within the order Rodentia, and has been known to live up to nine times as long as the average similarly-sized mouse (Buffenstein 2008).

Studies on several sirtuin deacetylases have suggested a link between acetylation and longevity within a variety of species. The first study to identify this link was in yeast, where cells mutant for the sirtuin *SIR2* did not show the normal increase in longevity seen with

caloric restriction (Lin *et al.* 2000). Looking at the sirtuins in mice, brain-specific overexpression of *Sirt1* increases lifespan (Satoh *et al.* 2013) and *Sirt6* overexpression has been shown to increase lifespan of male mice (Kanfi *et al.* 2012). Caloric restriction also alters acetylation in mouse liver mitochondria (Schwer *et al.* 2009). It is therefore interesting to consider the possible role acetylation plays in the differences in longevity observed between species.

Chapter summary

In this chapter, I describe the creation of a dataset of cysteine-lysine (Cys-Lys) and serine-lysine (Ser-Lys) pair conservation across a variety of vertebrate species. I analyse the conservation of pairs in different compartments (cytosol and mitochondrial matrix) and the driving cause where pairs are not conserved. I then investigate association of Cys-Lys pair conservation with lifespan.

Methods

Identifying close, surface pairs

Creation of 3D protein models and distance/surface calculations were carried out by Dr Anthony Smith of the Bioinformatics group, Medical Research Council Mitochondrial Biology Unit, University of Cambridge, CB2 0XY, UK. UniProt sequences of acetylated mouse proteins from Weinert *et al.* (2015) were aligned with a non-redundant PDB sequence database clustered at 95%, using MODELLER (Webb & Sali 2017). For proteins with a best match with over 50% sequence identity, the sequence was aligned with the structural template and five structures were predicted by using MODELLER. The structure with the lowest DOPE (discrete optimised protein energy) score was used for further calculations. Distance between atoms in the structure were calculated using trigonometric calculations and the solvent accessible area calculated using the program areaimol (Winn *et al.* 2011).

Identifying orthologues of acetylated mouse proteins

For mouse proteins with a calculated structural model, one-to-one human orthologues were identified where possible from Ensembl Compara (Herrero *et al.* 2016). Proteins without one-to-one human orthologues were removed from the analysis, to remove complication from paralogues. Orthologues of the human sequence in 70 vertebrate species (*Appendix IV – Table 1*) for each remaining protein were identified using the reciprocal best hit method (*Figure 7.2*). For each sequence, a BLASTp search (Altschul *et al.* 1990, 1997) was run against a local BLASTp database including non-identical protein sequences from the 70 different vertebrate species (E-value cut-off 1×10^{-10}). A second (reciprocal) BLASTp search was run against a BLASTp database of human proteins (E-value cut-off 1×10^{-10}), using the top hit from the original search for each species as the query sequence. The sequence was identified as an orthologue if the reciprocal BLASTp search returned a protein equivalent to the original human protein as the top hit.

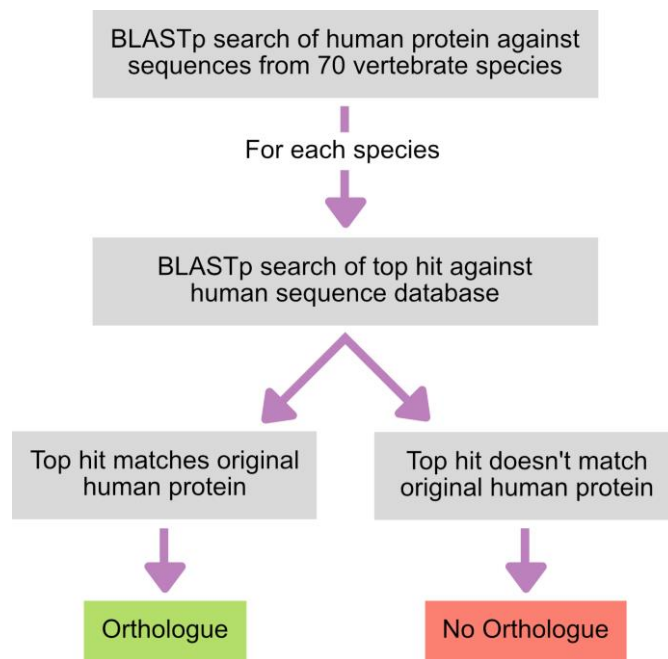


Figure 7.2 Summary of the reciprocal best hit method for identifying orthologues of proteins which contain detected acetylated residues in a mouse dataset from Weinert et al. (2015).

Proteins with fewer than sixty identified orthologues were removed from further analysis (Figure 7.3(a)). Five species with outlying numbers of orthologues were also removed from further analysis (Figure 7.3(b)).

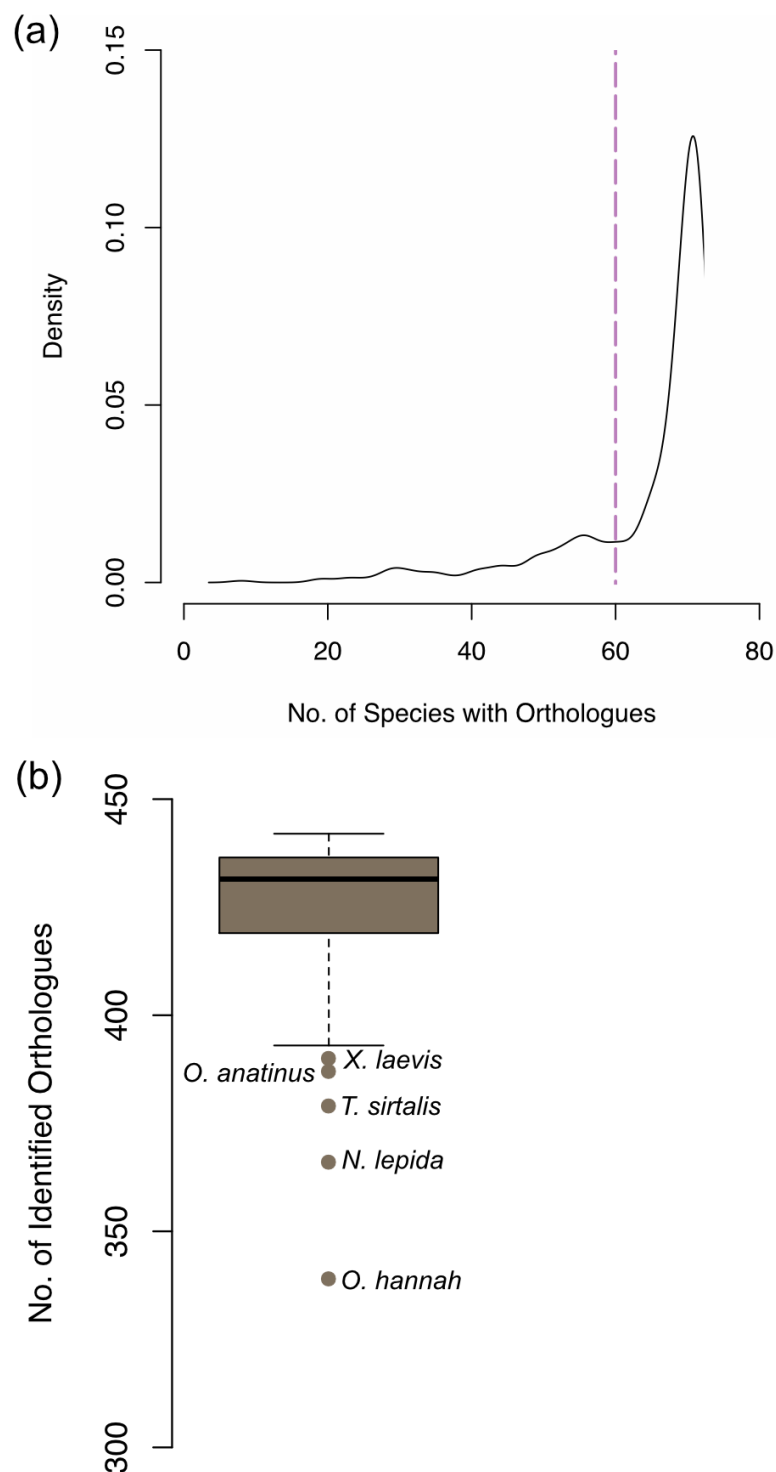


Figure 7.3 Quality control of the analysed proteins and species. (a) Density plot of the number of predicted orthologues for each targeted human protein. Dashed line is the cut-off for taking proteins forward for further analysis. (b) Boxplot of the number of detected orthologues for each included species. Labelled outliers (over $1.5 \times IQR + Q3$) were removed from further analysis.

Estimating relative conservation of residues and pairs

Sequences of predicted orthologues of each protein and the original mouse protein sequence were aligned using MUSCLE with default settings (Edgar 2004a, 2004b). Proximal, surface Cys-Lys and Ser-Lys pairs were identified in the predicted mouse structure, as those with $\leq 11.5 \text{ \AA}$ between the residues and both residues with $\geq 5 \text{ \AA}^2$ solvent exposure. Corresponding residues were identified for each species with an identified orthologue from the alignments with the mouse protein (*Figure 7.4*).

Pairs were assigned as either:

- Conserved – both residues match to the mouse pair,
- Not conserved – either one or both residues do not match to the mouse pair,
- Not present – either one or both positions has no assigned residue. This could be from inaccuracies in a predicted protein; a match to a different isoform or protein; or a real lack of residues at these positions.

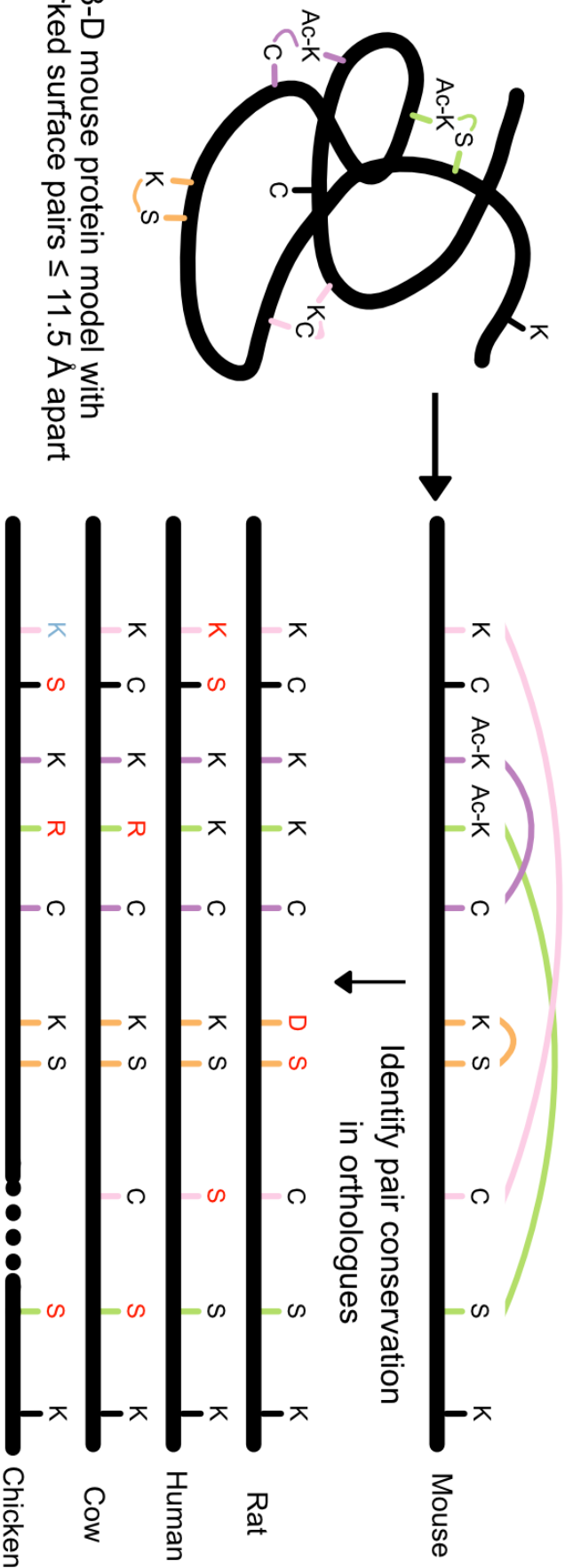


Figure 7.4 Measuring conservation of close, surface Cys-Lys and Ser-Lys pairs. Pairs of residues on the surface of mouse protein models, within 11.5 Å are marked in colour (green, pink, orange and purple). Orthologues from different species were aligned with the mouse protein, and the residues corresponding to the paired mouse residues were identified. Red residues indicate loss of a residue pair, due to change in one or both residues. Blue residues indicate lack of alignment in the second of the residue pair – this pair would be marked not present rather than not conserved.

Defining cytosolic and mitochondrial matrix proteins

Definitions of cytosolic and mitochondrial proteins were decided in collaboration with Dr Anthony Smith of the Bioinformatics group and Dr Andrew James of the Mitochondrial Dysfunction group, Medical Research Council Mitochondrial Biology Unit, University of Cambridge, CB2 0XY, UK. Proteins were designated as cytosolic if they were annotated in human or mouse Gene Ontology as cytosolic, or they were assigned as cytosolic with an evidence level of validated/supported in the Human Protein Atlas (Thul *et al.* 2017). Proteins were designated as mitochondrial matrix if they were annotated in human or mouse Gene Ontology as mitochondrial matrix, or identified as mitochondrial in an ascorbate peroxidase (APEX)-tagging study (Rhee *et al.* 2013), or assigned to the matrix in MitoCore – a model of mitochondrial metabolism (Smith *et al.* 2017). Proteins designated as both cytosolic and mitochondrial matrix were removed, to remove inaccuracies from differing levels of exposure to the two different chemical environments.

Estimating lifespan of sequenced vertebrates

Analysis of lifespan was limited to vertebrates, with particular effort to include a wide range of rodents, bats and birds. Included species had to have:

- a full sequenced genome in the NCBI nr database with protein annotation,
- a maximum lifespan listed in the AnAge Database (De Magalhães & Costa 2009; Tacutu *et al.* 2013),
- and an estimated adult bodyweight in grams, either from the AnAge Database or another recorded source.

Avian species included were either Sanger sequenced or noted as high-coverage genomes in G. Zhang *et al.* 2014.

To account for the relationship between body weight and lifespan, the maximum longevity (t_{max}) residual was calculated as described in the AnAge Database (Tacutu *et al.* 2013):

$$t_{max} \text{ residual} = \frac{\text{maximum observed lifespan (years)}}{4.88 \times \text{average adult weight (g)}^{0.153}} \times 100$$

Statistics

Statistical analyses were carried out in R. Statistical differences between residue groups was estimated using paired ANOVA and paired t -tests with Bonferroni correction. Differences in lifespan of mammal groups were tested using one-way ANOVA and post-hoc Tukey honest significant difference tests. Correlation was estimated as the Pearson product correlation coefficient (r) with an associated p -value.

Results

Creating a dataset of cysteine-lysine pair conservation

The degree to which acetylation is an enzymatically controlled regulatory process or a potentially damaging non-enzymatic reaction is still under discussion (Wagner & Hirschey 2014), particularly with the discovery that close cysteine-lysine pairs increase non-enzymatic $S \Rightarrow N$ lysine acetylation (James *et al.* 2017). One way to investigate functionality of residues is to look at the conservation across species – residues with important functional roles are more likely to be conserved across species.

Previous work has studied acetylation motifs in primary sequences (Rardin *et al.* 2013; Svinkina *et al.* 2015), but this analysis lacks information about the residues which are close in real protein structures due to tertiary structure. To address this, 3D structures were created of 619 mouse proteins with quantified acetylation (Weinert *et al.* 2015) and a strong homologous known structure (over 50% sequence identity). Calculations measuring the distance between the amino nitrogen of lysine residues and the thiol sulphur of cysteine residues allowed identification of close cysteine-lysine (Cys-Lys) pairs; and only solvent accessible ($> 5\text{\AA}^2$) pairs were taken forward, as non-enzymatic $S \Rightarrow N$ acetylation requires exposure of the residues to the environment. This produced a dataset of 52,321 distances between surface cysteines and lysines.

To investigate conservation, I used the reciprocal best hit method to identify orthologues of modelled proteins, across 72 vertebrate species. After quality control of species and proteins, I retained 29,173 orthologues of 442 proteins across 67 species (including mouse) for further analysis. Within these proteins, there are 890 surface Cys-Lys pairs and 6,353 surface Ser-Lys pairs located $\leq 11.5\text{\AA}$ apart in the predicted mouse protein structures. By aligning the identified orthologues, I identified the presence or absence of these pairs in the other studied species, creating a large dataset of pair conservation in comparison to mouse.

If the bulk of non-enzymatic $S \Rightarrow N$ acetylation is harmful, it may be expected that Cys-Lys pairs which are close enough to support non-enzymatic acetylation would be less conserved across species than non-acetylated Cys-Lys pairs. This was demonstrated in this dataset (*Figure 7.5(a)*) – across the tested species, Cys-Lys pairs which had observed *N*-acetylated lysines in mouse are less conserved than those located in the same proteins without observed

N-acetylation in mouse, (mean fraction conserved \pm standard deviation (s.d.); 0.63 ± 0.11 compared to 0.76 ± 0.09). Additionally, Cys-Lys pairs in the mitochondrial matrix (where conditions favour non-enzymatic acetylation) are significantly less conserved, on average, than Cys-Lys pairs in the cytosol (*Figure 7.5(b)*; mean fraction conserved \pm s.d.; 0.72 ± 0.09 compared to 0.76 ± 0.09). This is evidence that non-enzymatic $S \Rightarrow N$ acetylation may have some harmful effects, which are selected against.

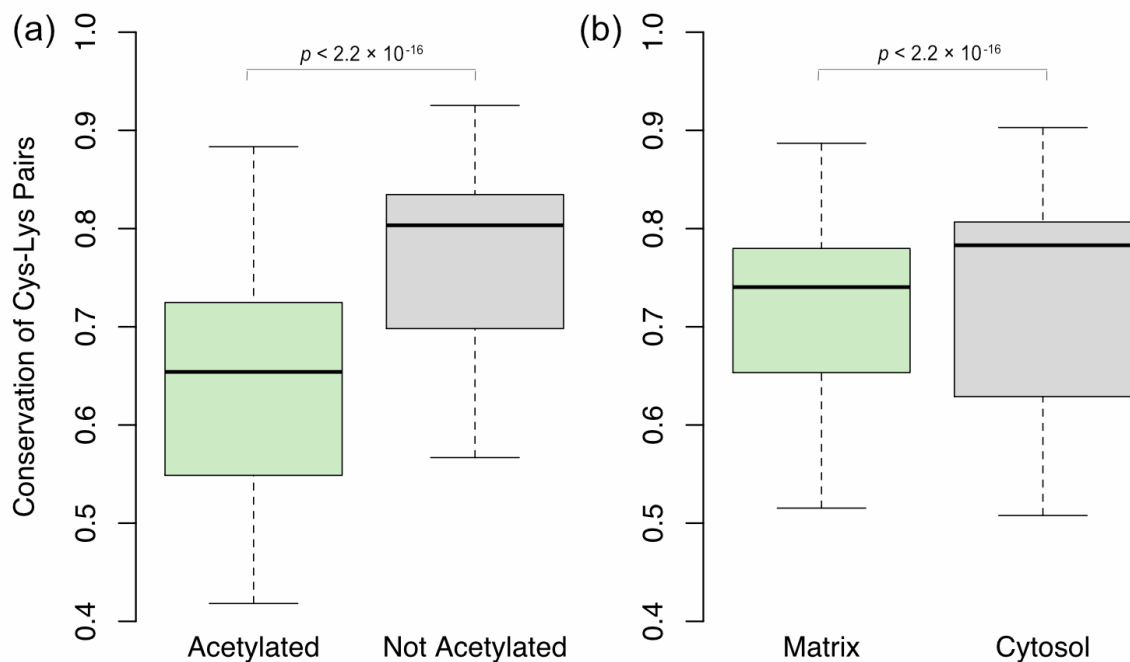


Figure 7.5 Fraction of conserved Cys-Lys pairs across the studied vertebrate species. (a) Separated by acetylation status of the lysine in mouse. (b) Separated by localisation of the protein. p-values from paired t-tests (paired by species).

Close acetylated Cys-Lys pairs are less conserved in the cytosol

The mitochondrial matrix contains high levels of acetyl-CoA and is more alkaline than the cytosol – a particularly good environment for non-enzymatic acetylation (Wagner & Payne 2013). One prediction from this is that there would be differences in the conservation of Cys-Lys pairs which are acetylated in mouse between the matrix and the cytosol, with the chemical environment providing different selective pressures.

Conservation would be expected to differ in species depending on evolutionary distance from mouse, as it is based on changes from observations in mouse. To correct for that bias, ratios of conservation of acetylated residue pairs to non-acetylated residue pairs were calculated for each species (where acetylation is based on the observed status in mouse). This gives a measure of relative conservation, independent of distance from mouse. Completely equal conservation would produce a ratio of 1, with higher ratios suggesting pairs acetylated in mouse were more conserved, and lower ratios that pairs acetylated in mouse were less conserved. Close, surface serine-lysine (Ser-Lys) pairs were used as a control, as serine is very close structurally to cysteine, but without the active sulphur necessary for non-enzymatic acetylation.

Conservation of both Cys-Lys and Ser-Lys pairs in the mitochondrial matrix are centred around one (equal conservation of putative acetylated and putative non-acetylated pairs) (*Figure 7.6(a)*). There is also no significant difference between the ratios for Cys-Lys pairs and Ser-Lys pairs in this cellular compartment. However, there is a significant difference between the ratios for Cys-Lys pairs and Ser-Lys pairs in the cytosol (*Figure 7.6(a)*; $p < 2.2 \times 10^{-16}$, paired t -test). Though the conservation ratios for both pair groups are centred lower than one, the Cys-Lys pair ratio is significantly lower than the Ser-Lys pair ratio – meaning putative acetylated pairs are less conserved than putative non-acetylated pairs. Putative acetylated Cys-Lys pairs in the cytosol are also less conserved than putative acetylated Ser-Lys pairs ($p = 2.8 \times 10^{-16}$, paired t -test), despite cysteines being the most highly conserved amino acid (Marino & Gladyshev 2010).

This analysis was repeated with only mammalian species, which are warm-blooded and maintain constant body temperature, as non-enzymatic $S \Rightarrow N$ acetylation is temperature dependent (James *et al.* 2017). Results were consistent within mammals (*Figure 7.6(b)*) – putative acetylated Cys-Lys pairs in the cytosol were less conserved than putative acetylated Ser-Lys pairs in the cytosol, but similar pairs in the matrix showed no difference in conservation.

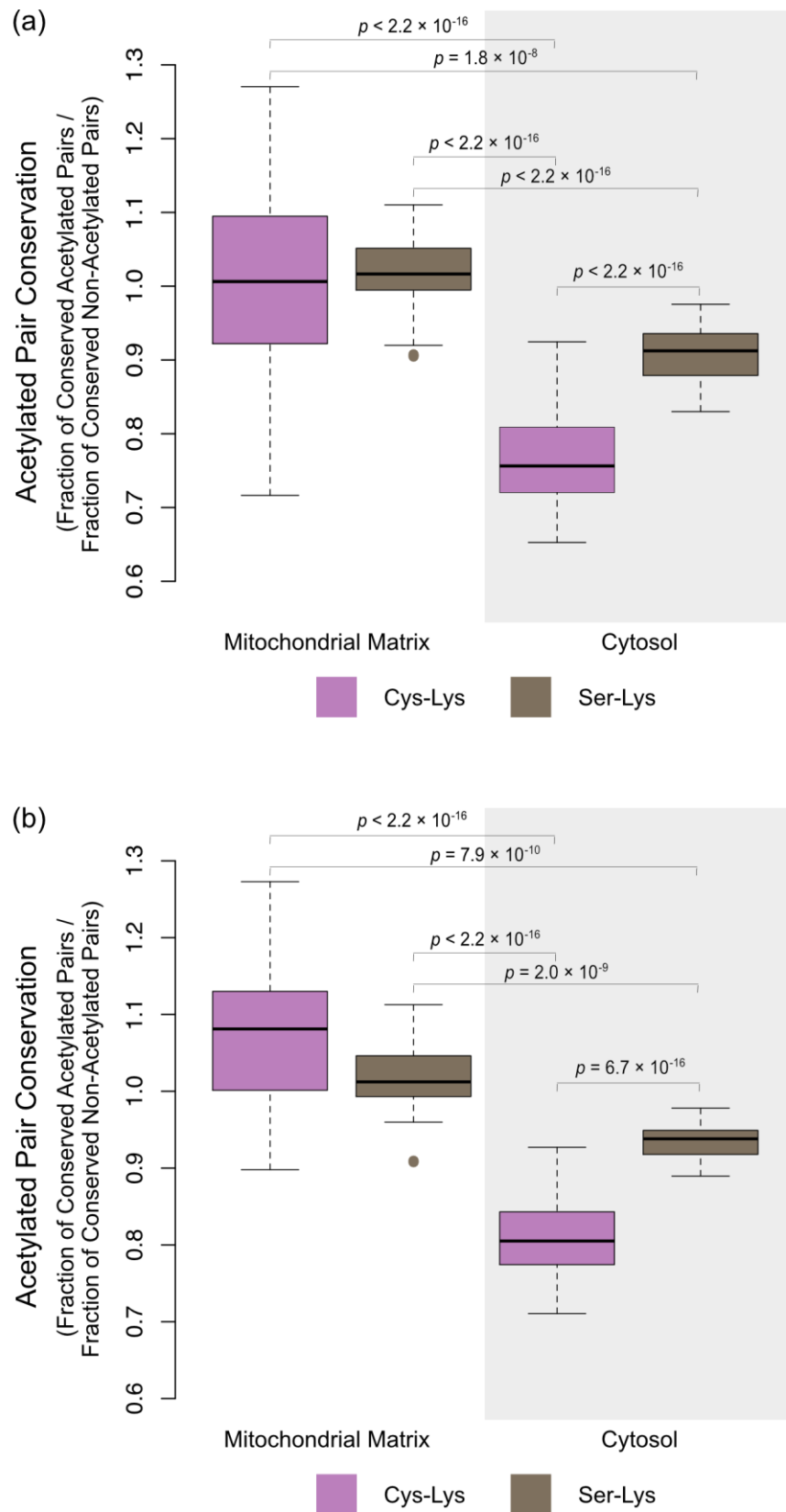


Figure 7.6 Proximal cytosolic Cys-Lys pairs that are known to be acetylated in mouse are less conserved. (a) All analysed species. (b) Only mammalian species. p-values reported from paired t-tests with Bonferroni multiple testing correction.

The lack of conservation of residue pairs could be driven by lack of conservation of either residue singly or both together. To analyse this, I looked at the conservation of single residues from cytosolic pairs known to be acetylated in mouse, compared to cytosolic residues not in acetylated pairs in mouse (*Figure 7.7*). The mean conservation ratios for lysine in Cys-Lys pairs (0.965), serine in Ser-Lys pairs (0.955) and lysine in Ser-Lys pairs (0.953) are all close to 1 and the means are not significantly different. However, the mean conservation ratio for cysteine in Cys-Lys pairs is lower at 0.836 and is significantly different from the other residue groups tested. The range of conservation is also higher for cysteines in Cys-Lys pairs – i.e. there is a greater spread in conservation ratios, which drives the spread in conservation of cytosolic Cys-Lys pairs.

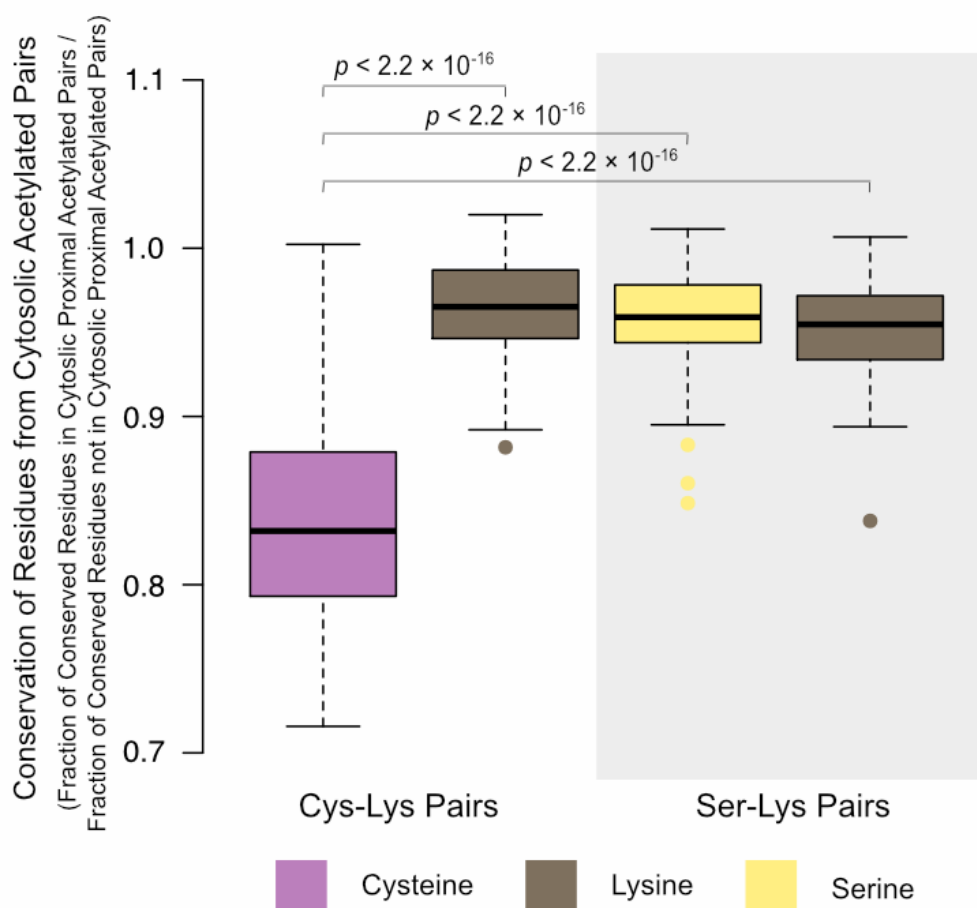


Figure 7.7 Changes in cysteine conservation drive changes in Cys-Lys pair conservation in the cytosol. p-values reported from paired t-tests with Bonferroni correction.

Conservation of cytosolic Cys-Lys pairs correlates with a measure of mammalian lifespan

Studies observing the effect of sirtuin deacetylases on lifespan in a variety of species, from yeast (Lin *et al.* 2000) to mice (Kanfi *et al.* 2012), provide a possible link between lifespan and acetylation. One way to investigate a possible effect of $S \Rightarrow N$ acetylation is to look at the conservation of Cys-Lys pairs which are acetylated in mouse across species – if $S \Rightarrow N$ acetylation has a detrimental effect on lifespan, putative acetylated Cys-Lys pairs may be less conserved in longer-lived species.

As the number of species included was relatively small compared to previous studies of lifespan, I first checked to see if the strong correlation previously found between higher adult weight and higher adult lifespan held (*Figure 7.8*). While individually only the birds and ‘other mammals’ groupings have significant correlation, the combined correlation for all included species is significant, ($p = 1.69 \times 10^{-9}$), with increasing size correlated with increasing lifespan ($r = 0.638$). This confirms that the known longevity trend is maintained even in this relatively small sample of species.

The bats included in this analysis show the opposite pattern to that expected (although the correlation is not statistically significant) – smaller bats are actually the longest lived. This appears to be mainly driven by the two smallest bats included (*Myotis lucifugus* and *Myotis brandtii*), which are two of the longest lived bat species known (Wilkinson & South 2002; Podlutsky *et al.* 2005).

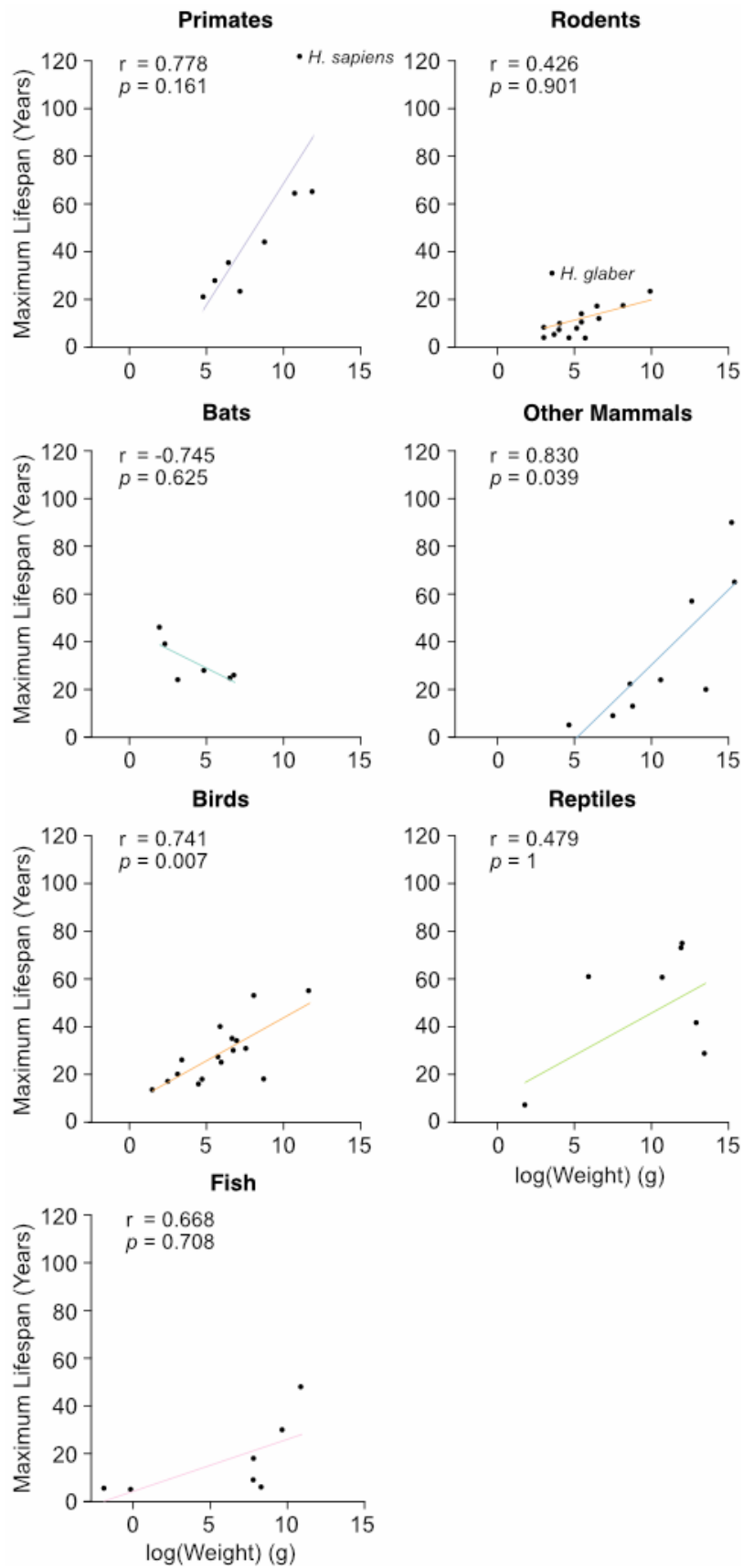


Figure 7.8 Lifespan correlates with adult body weight in vertebrates. Pearson correlation coefficient (r) and p-value reported for each grouping of species.

Another important previous observation was that bats tend to live longer than non-flying mammals of a similar weight (Austad & Fischer 1991). With only a small sample of bats to work from, it was important to check that this finding held in this sample analysis. To do this, I used a longevity measure (maximum longevity (t_{\max}) residual) which compares the maximum observed lifespan to a predicted lifespan based on weight (Tacutu *et al.* 2013). A t_{\max} residual higher than 100% reflects a species with a higher observed lifespan than predicted based on adult size; whilst if a species has a t_{\max} residual below 100% it has a lower observed lifespan than predicted based on adult size. Only mammals were included as this measure is based on the mammalian allometric equation. The mammals were split into four groupings: primates, rodents, bats and other mammals.

ANOVA analysis identified a significant difference between the groups ($p < 2.2 \times 10^{-16}$). Bats are significantly longer-lived than rodents ($p = 0.001$) and other mammals ($p = 0.003$), despite the small sample size of only six bats (*Figure 7.9*). Two labelled outliers are humans (*Homo sapiens*) in the primates and the naked mole-rat (*Heterocephalus glaber*) in the rodents. The naked mole-rat is known for being an exceptionally long-lived rodent (Sherman & Jarvis 2002), whilst humans are the best observed species with many methods of extending lifespan (through medicine etc.).

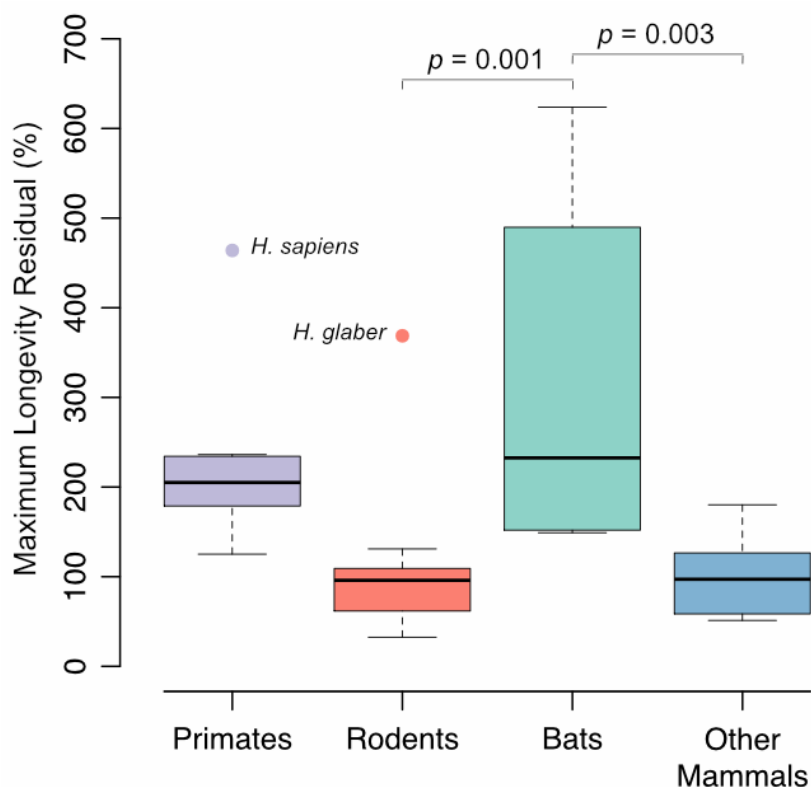


Figure 7.9 Bats live significantly longer for their weight than rodents and the group ‘other mammals’. Maximum longevity residual is a measure of percent of predicted lifespan for maximum observed lifespan, corrected for adult body weight. p-values from Tukey honest significant difference post-hoc analysis.

I then plotted mammalian lifespan, as measured by the t_{\max} residual, against conservation of cytosolic Cys-Lys pairs which are acetylated in mouse for these species. t_{\max} residual is significantly and highly correlated with the conservation of cytosolic putative acetylated Cys-Lys pairs ($r = -0.750$, $p = 1.36 \times 10^{-7}$), but not cytosolic putative acetylated Ser-Lys pairs ($r = -0.285$, $p = 0.092$) (Figure 7.10(a)). Maximum lifespan, not corrected for bodyweight, is also correlated with conservation of cytosolic putative acetylated Cys-Lys pairs, but slightly less strongly ($r = -0.531$, $p = 0.0008$). The same pattern is not seen for mitochondrial matrix putative acetylated Cys-Lys pairs ($r = -0.129$, $p = 0.356$), though there is a weak but significant correlation with mitochondrial matrix acetylated Ser-Lys pairs ($r = 0.380$, $p = 0.022$) (Figure 7.10(b)).

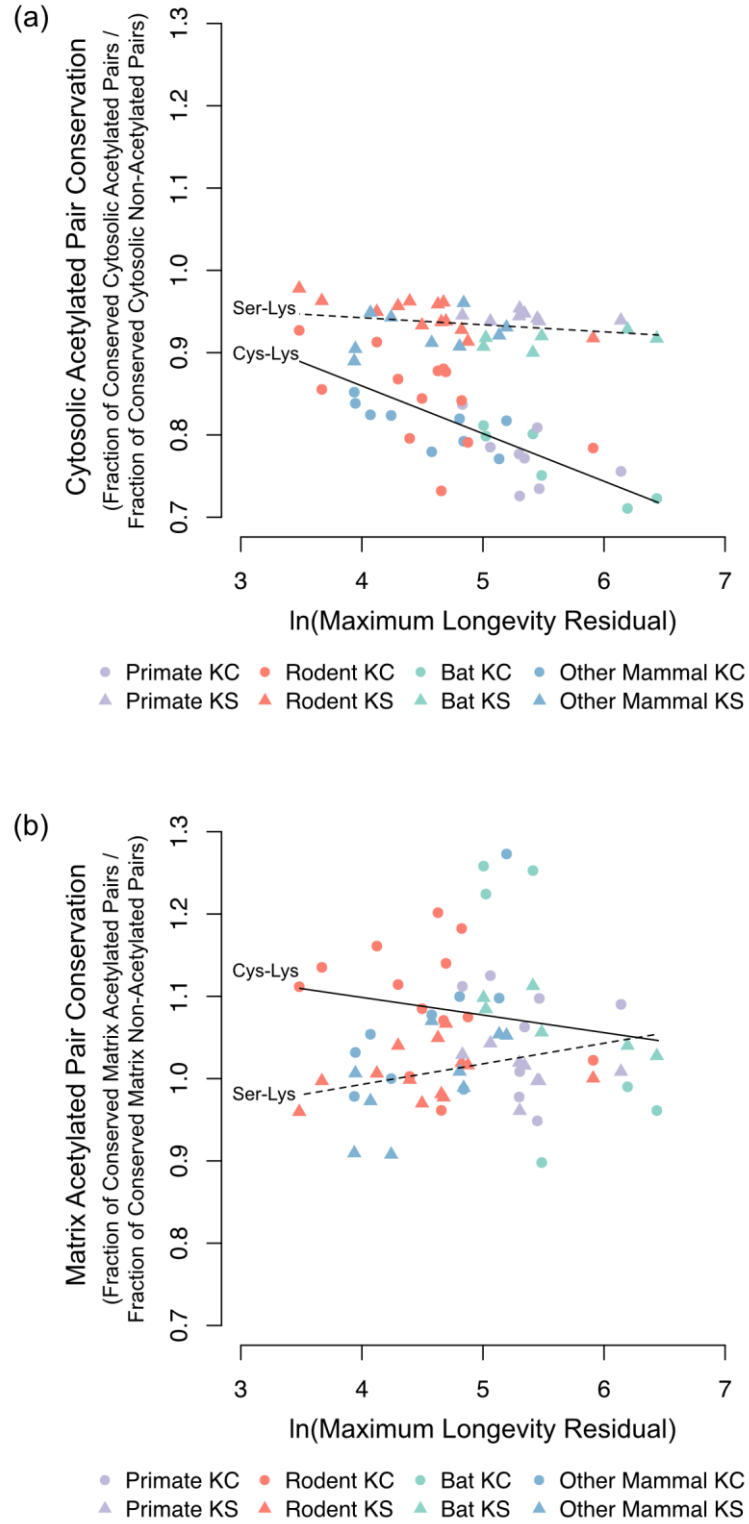


Figure 7.10 Correlation of the conservation of putative acetylated Cys-Lys (KC) and Ser-Lys (KS) pairs with a measure of mammalian lifespan – the maximum longevity residual. (a) Cytosolic proteins. For KC pairs: $r = -0.750$, $p = 1.36 \times 10^{-7}$. For KS pairs: $r = -0.285$, $p = 0.092$. (b) Proteins localised to the mitochondrial matrix. For KC pairs: $r = -0.129$, $p = 0.356$. For KS pairs: $r = -0.380$, $p = 0.022$.

Discussion

Previous work had shown that non-enzymatic $S \Rightarrow N$ lysine acetylation with transfer of an acetyl group from acetyl-CoA via a cysteine residue occurs, at least *in vitro* (James *et al.* 2017). This had not been demonstrated *in vivo* and the role of the majority of acetylation is still under discussion. As conservation can be used to investigate function, I identified vertebrate orthologues of modelled mouse proteins known to have acetylated lysine residues (Weinert *et al.* 2015) and used these to estimate conservation of 890 surface Cys-Lys pairs less than 11.5Å apart.

In this dataset, Cys-Lys pairs which are acetylated in mouse are significantly less conserved than non-acetylated Cys-Lys pairs (*Figure 7.5*), suggesting non-enzymatic $S \Rightarrow N$ acetylation may be damaging and, thus, selected against. This is despite acetylated lysines being more conserved than non-acetylated lysines overall (Henriksen *et al.* 2012), with acetylation being well conserved across species (Weinert *et al.* 2011; Rardin *et al.* 2013). Additional analyses supporting the possible *in vivo* damaging effect of non-enzymatic $S \Rightarrow N$ acetylation using the modelled protein dataset were carried out by Dr Andrew James of the Mitochondrial Dysfunction group, Medical Research Council Mitochondrial Biology Unit, University of Cambridge, CB2 0XY, UK. He found that close surface Cys-Lys pairs are significantly less frequent in the acetylating environment of the mitochondrial matrix than the cytosol; that cysteine residues are significantly less exposed to the protein surface in the matrix than in the cytosol; and that Cys-Lys pairs are significantly further apart than corresponding Ser-Lys pairs in both the mitochondrial matrix (0.47Å difference) and the cytosol (0.29Å difference). Damaging effects of having close surface Cys-Lys pairs due to non-enzymatic $S \Rightarrow N$ acetylation may contribute to the lower conservation observed for cysteines on the surface of proteins compared to internal cysteines (Marino & Gladyshev 2010).

Though the frequency of surface Cys-Lys pairs in the mitochondrial matrix is lower than in the cytosol, Cys-Lys pairs which are acetylated in mouse are not significantly less conserved than Ser-Lys pairs which are acetylated in mouse, in this compartment (*Figure 7.6*). However, there is a significant difference in conservation between putative acetylated Cys-Lys pairs and putative acetylated Ser-Lys pairs in the less favourable acetylating environment of the cytosol (*Figure 7.6*), driven by decreased conservation of cysteine residues close to putative acetylated lysines (*Figure 7.7*). One possibility is that there has already been strong selection

against the close Cys-Lys pairs in the mitochondria, earlier in the eukaryotic phylogenetic tree than the branching of the studied vertebrates, which has maximised non-functional pair loss in this environment. This could be investigated by looking at conservation of Cys-Lys pairs in mitochondrial matrix proteins beyond vertebrates, though increasingly differing lifestyles and environmental conditions may also influence changes in amino acid frequency or acetylation profiles. The less favourable environment in the cytoplasm (lower pH and lower concentration of acetyl-CoA) may have decreased selective pressure for the loss of Cys-Lys pairs.

Conservation of cytosolic putative acetylated Cys-Lys pairs, but not Ser-Lys pairs or matrix Cys-Lys pairs, correlates strongly with maximum lifespan and t_{\max} residual, which corrects for differences in body weight (*Figure 7.10*). This contributes to the idea that non-enzymatic $S \Rightarrow N$ acetylation may be damaging, as species that live longer have less of these putative acetylated pairs. It was unexpected that this link was identified in cytosolic proteins rather than matrix proteins, because of the respective chemical environments of the two compartments. However, though the matrix is the highly acetylating environment, a non-mitochondrial sirtuin deacetylase was actually first associated with changes in lifespan (Lin *et al.* 2000) and other non-mitochondrial sirtuin deacetylases have since also been associated with changes in lifespan (Kanfi *et al.* 2012; Satoh *et al.* 2013).

Though the link between non-enzymatic $S \Rightarrow N$ acetylation and possible cell damage is consistent with the very low stoichiometry of most quantified acetylation (Weinert *et al.* 2015), only a minority (5.7%) of *N*-acetylated lysine residues in this study have a surface cysteine residue within 11.5Å, so other mechanisms of acetylation are clearly important. This could be enzymatic acetylation or other forms of non-enzymatic acetylation. This analysis was based on *N*-acetylation identified in mouse liver tissue, but conditions in different tissues (including concentrations of deacetylases or other molecules that interfere with $S \Rightarrow N$ acetyl transfer, such as glutathione) may change the magnitude of contribution from non-enzymatic $S \Rightarrow N$ acetylation (James *et al.* 2017).

The cellular changes caused by uncontrolled, non-enzymatic $S \Rightarrow N$ acetylation and, therefore, the possible effects on fitness which allow selection against close Cys-Lys pairs are unclear. *N*-acetylation of lysine residues removes the positive charge and sterically blocks the lysine residue from further reactivity (Choudhary *et al.* 2014). It may affect protein interaction with other proteins, cofactors or substrates, and/or increase protein aggregation

(Kuczyńska-Wiśnik *et al.* 2016). Differences between how these problems are dealt with in the mitochondria and the cytoplasm may also contribute to the magnitude of the problem caused by $S \Rightarrow N$ acetylation in these compartments.

Possible damaging effects of non-enzymatic $S \Rightarrow N$ acetylation may also help explain the link between changes in acetylation and some disease states. For example, *SIRT1* deacetylase overexpression has been shown to provide some protection against neuronal death in a mouse model of Alzheimer's disease (Kim *et al.* 2007) and increased acetylation caused by decreased *Sirt3* activity (due to a reduced NAD^+/NADH ratio from dysfunction of complex I) has been linked to heart failure (Lee & Tian 2015).

Conclusions

By studying the conservation of structurally close Cys-Lys pairs on the surface of proteins, I provide evidence that non-enzymatic $S \Rightarrow N$ acetylation is likely to be damaging to cells. Though the mitochondrial matrix is the more acetylating environment, only cytosolic putative acetylated Cys-Lys pairs show changes in conservation in comparison to non-acetylated pairs. This decrease in conservation is driven by changes in cysteine residues. Conservation of cytosolic putative acetylated Cys-Lys pairs is correlated with mammalian lifespan corrected for bodyweight, providing a possible link between acetylation and ageing.

Chapter 8

Conclusions

Thesis summary

The general aim of this thesis was to use the new definition of the human mitochondrial proteome, provided by IMPI 2017, as a basis to explore the phylogenetic history of the human mitochondrial proteome, through the identification of orthologues. This orthology dataset was then used in the discovery of new knowledge about mitochondrial disease and the function of the mitochondrial proteome.

The first step was to create the mitochondrial proteome orthology dataset, by identifying orthologues of IMPI 2017 genes across a range of species (both eukaryotic and prokaryotic), as described in *Chapter 2*. Reciprocal best hit analysis, followed by a manual curation process, was used to produce a large dataset of 190,097 predicted orthologous sequences of 1,550 genes across 359 species. Nearly three quarters of the genes had predicted orthologues only in eukaryotic species and nearly half only had predicted orthologues in metazoan species. This reflects a huge increase in mitochondrial complexity since the mitochondrion's roots in the α -proteobacteria, just as other parts of the eukaryotic cell have gained in complexity.

The orthology dataset provided a solid starting point for investigations into the function of the uncharacterised parts of the mitochondrial proteome. Phylogenetic profiling relies on the identification of a pattern of presence and absence of genes across a range of species to predict the potential functions of an uncharacterised gene, under the assumption that genes involved in similar structures or pathways will have similar profiles. I carried out phylogenetic profiling analyses of the respiratory complexes, electron transfer flavoprotein and ATP synthase, by using the orthology dataset to create phylogenetic profiles of all the IMPI 2017 genes (*Chapter 3*). There were no new predictions for genes associated with complexes I-IV or ATP synthase, though *C8orf82* is a good candidate for a gene associated with the function of the electron transfer flavoprotein.

By considering the results of phylogenetic profiling analyses, I derived a set of guidelines to increase the probability of success for future phylogenetic analysis in eukaryotes. To increase the chance of success, complexes and pathways studied by phylogenetic profiling should show good variation in their presence and absence across the studied species. When looking for the function of human proteins, it is useful to have variation in species which share a common ancestor with humans as recently as possible. This allows the inclusion in the

analysis of a higher number of recent gene innovations, as well as gene loss events in parasitic species. Having a set of closely related species within a single taxon, with variation in the phylogenetic profile of the studied complex/pathway, is also useful, as exemplified by the differences in complex I presence in closely related yeasts.

Using these guidelines, I designed a further phylogenetic profiling analysis to identify potential transport substrates for uncharacterised members of the mitochondrial carrier family, as described in *Chapter 4*. Variation in the presence and absence of several mitochondrial carriers was observed in two metazoan phyla (the nematodes and the platyhelminthes). A dataset of predicted orthologues of genes encoding the mitochondrial proteome in a range of sequenced species from each taxon was, therefore, created to facilitate profiling analysis. This was first used to study the characterised mitochondrial carriers, as a proof that this approach could produce sensible results. Phylogenetic profiling produced a good prediction for the transport substrate of one carrier (*SLC25A21*) in the platyhelminthes and provided less obvious links to the transport substrates of a few other carriers. However, these predictions were not consistent when comparing the results from the two phyla, so in further analysis the results from the two phyla were considered separately.

The knowledge gained from investigation of the characterised carriers was used to structure the study of four groups of uncharacterised carriers which had different patterns of presence and absence within the two studied phyla. Potential transport substrates were predicted for two of the four groups, based on the phylogenetic profile and literature evidence. These were ketone bodies for the *SLC25A14/30* carriers and cyclic pyranopterin monophosphate (a precursor for molybdenum cofactor) for the *SLC25A44* carrier. This evidence provides a good case for wet laboratory work to test the function of these two carrier groups and testing for the transport of ketone bodies by *SLC25A14/30* has been started. This work shows the potential of phylogenetic profiling to identify eukaryotic gene functions when the target genes and species studied are chosen with care.

A second use of the original orthology dataset described in *Chapter 2* was to investigate phylogenetic features of mitochondrial genes associated with monogenetic disease (*Chapter 5*). Disease genes had older predicted orthologues, as well as a wider spread of orthologues in different taxa, than genes that are not currently associated with disease. However, the opposite was true of genes associated with dominantly inherited diseases which were younger, with less phylogenetic spread, than genes associated with recessive disease.

These features have been included in a machine learning approach to predict mitochondrial disease genes, which will be used as part of a pipeline to prioritise candidate disease genes in the sequenced exome/genome of patients with mitochondrial disease.

After exploring the orthology of genes encoding the mitochondrial proteome in cellular organisms, I was inspired to investigate the same in the other possible domain of life – the viruses (*Chapter 6*). A wide range of viruses had predicted orthologues of at least one gene associated with the mitochondrial proteome, and these genes were often associated with processes that could be linked to viral replication (e.g. DNA replication, translation). Despite this wide range, most genes were only identified in a single or few viral families.

The final analysis dealt with in this thesis investigated the conservation of sites of potential non-enzymatic $S \Rightarrow N$ acetylation, across vertebrate species (*Chapter 7*), to explore the hypothesis that these pairs are potentially damaging. Acetylated and close cysteine-lysine pairs on the surface of proteins were less conserved than non-acetylated pairs, but only in cytosolic proteins (not proteins in the mitochondrial matrix). The level of conservation of these pairs correlates with mammalian lifespan – species with less conservation of these pairs have longer lifespan adjusted for bodyweight. This supports the hypothesis that non-enzymatic $S \Rightarrow N$ acetylation may be potentially damaging, at least in the cytosol; enough so that these pairs are less conserved in longer lived species.

Further work

The creation of an extensive orthology dataset of human mitochondrial proteins provides a starting point for many potential additional analyses. The difficulty in assigning some orthologues from related groups of genes and the lack of literature for many of these groups suggests that there is still much work to be done in the clarification of the history of paralogous gene groups. Additionally, most of the work described in this thesis has been on a gene level – looking at the predicted presence and absence of genes. However, the orthology dataset consists of the actual sequences identified as predicted orthologues, facilitating sequence level comparisons across species.

One example of a potential sequence level analysis could be the examination of the performance of different mitochondrial targeting sequence (MTS) prediction programs across

different taxa. There are a variety of available MTS prediction programs which use different methodologies to attempt to predict *N*-terminal MTSs from a protein sequence (Sun & Habermann 2017). The sequences from the orthology dataset could be used to compare the performance of these programs in different taxa, as it may be the case that certain approaches work better in some species. Preliminary work in generating IMPI 2017 has shown the recognition of human mitochondrially-targeted proteins by machine learning is improved by considering whether predicted MTSs are conserved among orthologues from mouse, rat and zebrafish.

Considering future phylogenetic profiling efforts, several additional taxa may be of particular interest, particularly as genomes become available for more obscure metazoan species. For example, in this analysis, the parasitic cnidarian *Thelohanellus kitauei* showed great variance in predicted mitochondrial metabolism compared to other cnidaria, as would be predicted by the huge metabolic losses previously described (Yang *et al.* 2014). Sequencing and analysis of other myxozoans will therefore be of interest. There has also been a controversial identification of metazoan species (members of the Loricifera) living in an anoxic environment (Danovaro *et al.* 2010), though this has been challenged (Mentel *et al.* 2016). Loriciferans may be a useful taxon for phylogenetic profiling once genomes from these species start to be sequenced.

My investigations of metabolism have also identified pathways which may be particularly amenable to phylogenetic profiling analysis, due to the identified variation in their presence and absence across species. One example is the metabolism of the cofactors derived from vitamin B12: methylcobalamin (synthesised in the cytoplasm in mammals) and adenosylcobalamin (synthesis in the mitochondria in mammals) (Zhang *et al.* 2009). Differences in the synthesis and utilisation of adenosylcobalamin were identified in the nematodes, platyhelminthes and members of the Pancrustacea studied. Phylogenetic study of potential transporters in these taxa could be of interest, as the import mechanism of vitamin B12 derivatives into the mitochondria is unresolved (Froese & Gravel 2010).

Final words

In summary, in this thesis I have used studies of the orthology and phylogeny of genes of the human mitochondrial proteome to explore several aspects of mitochondrial function, disease and ageing. Part of this thesis has informed the start of targeted functional testing for some uncharacterised mitochondrial carriers. Information on the phylogenetic spread and history of genes is being included in a machine learning approach to identify candidate genes from sequence data of mitochondria disease patients. The mitochondrial proteome orthology dataset produced by this work provides opportunities for many further analyses of mitochondrial function and history.

References

- Abad, M.F.C., di Benedetto, G., Magalhães, P.J., Filippin, L., and Pozzan, T., 2004. Mitochondrial pH monitored by a new engineered green fluorescent protein mutant. *The Journal of Biological Chemistry*, **279**(12), pp.11521-11529.
- Abbasi, A.A., 2010. Unraveling ancient segmental duplication events in human genome by phylogenetic analysis of multigene families residing on HOX-cluster paralogs. *Molecular Phylogenetics and Evolution*, **57**(2), pp.836-848.
- Adly, N., Alhashem, A., Ammari, A., and Alkuraya, F.S., 2014. Ciliary genes *TBC1D32/C6orf170* and *SCLT1* are mutated in patients with OFD type IX. *Human Mutation*, **35**(1), pp.36-40.
- Ahn, C.S. and Metallo, C.M., 2015. Mitochondria as biosynthetic factories for cancer proliferation. *Cancer & Metabolism*, **3**(1), pp.1-10.
- Ahrens-Nicklas, R.C., Umanah, G.K.E., Sondheimer, N., Deardorff, M.A., Wilkens, A.B., Conlin, L.K., Santani, A.B., Nesbitt, A., Juulsola, J., Ma, E., Dawson, T.M., Dawson, V.L., and Marsh, E.D., 2017. Precision therapy for a new disorder of AMPA receptor recycling due to mutations in *ATAD1*. *Neurology Genetics*, **3**(1), e130.
- Akiyoshi, D.E., Morrison, H.G., Lei, S., Feng, X., Zhang, Q., Corradi, N., Mayanja, H., Tumwine, J.K., Keeling, P.J., Weiss, L.M., and Tzipori, S., 2009. Genomic survey of the non-cultivable opportunistic human pathogen, *Enterocytozoon bieneusi*. *PLoS Pathogens*, **5**(1), e1000261.
- Altenhoff, A.M. and Dessimoz, C., 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, **5**(1), e1000262.
- Altenhoff, A.M., Studer, R.A, Robinson-Rechavi, M., and Dessimoz, C., 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Computational Biology*, **8**(5), e1002514.
- Altenhoff, A.M., *et al.*, 2016. Standardized benchmarking in the quest for orthologs. *Nature Methods*, **13**(5), pp.425-430.

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), pp.403-410.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), pp.338-3402.
- Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A., 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, **43**(D1), pp.D789-D798.
- Amoedo, N.D., Punzi, G., Obre, E., Lacombe, D., de Grassi, A., Pierri, C.L., and Rossignol, R., 2016. *AGC1/2*, the mitochondrial aspartate-glutamate carriers. *Biochimica et Biophysica Acta*, **1863**(10), pp.2394-2412.
- Anderson, K.A. and Hirschey, M.D., 2012. Mitochondrial protein acetylation regulates metabolism. *Essays in Biochemistry*, **52**, pp.23-35.
- Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J.H., Staden, R., and Young, I.G., 1981. Sequence and organization of the human mitochondrial genome. *Nature*, **290**(5806), pp.457-465.
- Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontén, T., Alsmark, U.C.M., Podowski, R.M., Näsland, A.K., Eriksson, A-S., Winkler, H.H., and Kurland, C.G., 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**(6707), pp.133-140.
- Andrews, B., Carroll, J., Ding, S., Fearnley, I.M., and Walker, J.E., 2013. Assembly factors for the membrane arm of human complex I. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(47), pp.18934-18939.
- Angebault, C., Charif, M., Guegen, N., Piro-Megy, C., Mousson de Camaret, B., Procaccio, V., Guichet, P-O., Hebrard, M., Manes, G., Leboucq, N., Rivier, F., Hamel, C.P., Lenaers, G., and Roubertie, A., 2015. Mutation in *NDUFA13/GRIM19* leads to early onset hypotonia, dyskinesia and sensorial deficiencies, and mitochondrial complex I instability. *Human Molecular Genetics*, **24**(14), pp.3948-3955.

- Aravind, L., Dixit, V.M., and Koonin, E.V., 2001. Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. *Science*, **291**(5507), pp.1279-1284.
- Ariza, M-E., Glaser, R., Kaumaya, P.T.P., Jones, C., and Williams, M.V., 2009. The EBV-encoded dUTPase activates NF-kappa B through the *TLR2* and *MyD88*-dependent signaling pathway. *Journal of Immunology*, **182**(2), pp.851-859.
- Arnoldi, A., Tonelli, A., Crippa, F., Villani, G., Pacelli, C., Sironi, M., Pozzoli, U., D'Angelo, M.G., Meola, G., Martinuzzi, A., Crimella, C., Redaelli, F., Panzeri, C., Renieri, A., Comi, G.P., Turconi, A.C., Bresolin, N., and Bassi, M.T., 2008. A clinical, genetic, and biochemical characterization of *SPG7* mutations in a large cohort of patients with hereditary spastic paraplegia. *Human Mutation*, **29**(4), pp.522–531.
- Aung, L.H.H., Li, R., Prabhakar, B.S., and Li, P., 2017. Knockdown of *Mtfp1* can minimize doxorubicin cardiotoxicity by inhibiting *Dnm1l*-mediated mitochondrial fission. *Journal of Cellular and Molecular Medicine*, **21**(12), pp.3394-3404.
- Aung, L.H.H., Li, R., Prabhakar, B.S., Maker, A.V., and Li, P., 2017. Mitochondrial protein 18 (*MTP18*) plays a pro-apoptotic role in chemotherapy-induced gastric cancer cell apoptosis. *Oncotarget*, **8**(34), pp.56582-56597.
- Austad, S.N. and Fischer, K.E., 1991. Mammalian aging, metabolism, and ecology: evidence from the bats and marsupials. *Journal of Gerontology*, **46**(2), pp.B47-B53.
- Baeza, J., Smallegan, M.J., and Denu, J.M., 2016. Mechanisms and dynamics of protein acetylation in mitochondria. *Trends in Biochemical Sciences*, **41**(3), pp.231-244.
- Baltimore, D., 1971. Expression of animal virus genomes. *Bacteriological Reviews*, **35**(3), pp.235-241.
- Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S., 2002. Extensive feature detection of *N*-terminal protein sorting signals. *Bioinformatics*, **18**(2), pp.298-305.

- Bie, A.S., Fernandez-Guerra, P., Birkler, R.I.D., Nisemblat, S., Pelnena, D., Lu, X., Deignan, J.L., Lee, H., Dorrani, N., Corydon, T.J., Palmfeldt, J., Bivina, L., Azem, A., Herman, K., and Bross, P., 2016. Effects of a mutation in the *HSPE1* gene encoding the mitochondrial co-chaperonin HSP10 and its potential association with a neurological and developmental disorder. *Frontiers in Molecular Biosciences*, **3**, p.65.
- Binder, J.X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue S.I., Schneider, R., and Jensen, L.J., 2014. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database*, **2014**, bau012.
- Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R., 2009. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, **25**(22), pp.3045-3046.
- Bisaccia, F., De Palma, A., and Palmieri, F., 1989. Identification and purification of the tricarboxylate carrier from rat liver mitochondria. *Biochimica et Biophysica Acta*, **977**(2), pp.171-176.
- Blekhman, R., Man, O., Herrmann, L., Boyko, A.R., Indap, A., Kosiol, C., Bustamante, C.D., Teshima, K.M., and Przeworski, M., 2008. Natural selection on genes that underlie human disease susceptibility. *Current Biology*, **18**(12), pp.883-889.
- Blueweiss, L., Fox, H., Kudzma, V., Nakashima, D., Peters, R., and Sams, S., 1978. Relationships between body size and some life history parameters. *Oecologia*, **37**(2), pp.257-272.
- Brooks, S.S., Wall, A.L., Golzio, C., Reid, D.W., Kondyles, A., Willer, J.R., Botti, C., Nicchitta, C.V., Katsanis, N., and Davis, E.E., 2014. A novel ribosomopathy caused by dysfunction of *RPL10* disrupts neurodevelopment and causes X-linked microcephaly in humans. *Genetics*, **198**(2), pp.723-733.
- Brynolf, K., Eliasson, R., and Reichard, P., 1978. Formation of Okazaki fragments in polyoma DNA synthesis caused by misincorporation of uracil. *Cell*, **13**(3), pp.573-580.
- Buffenstein, R., 2008. Negligible senescence in the longest living rodent, the naked mole-rat: insights from a successfully aging species. *Journal of Comparative Physiology: B Biochemical, Systemic, and Environmental Physiology*, **178**(4), pp.439-445.

- Bych, K., Kerscher, S., Netz, D.J.A, Pierik, A.J., Zwicker, K., Huynen, M.A., Lill, R., Brandt, R., and Balk, J., 2008. The iron-sulphur protein *Ind1* is required for effective complex I assembly. *The EMBO Journal*, **27**(12), pp.1736-1746.
- Cai, J.J., Borenstein, E., Chen, R., and Petrov, D.A., 2009. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. *Genome Biology and Evolution*, **1**, pp.131-144.
- Calvo, S.E., Compton, A.G., Hershman, S.G., Lim, S.C., Lieber, D.S., Tucker, E.J., Laskowski, A., Garone, C., Liu, S., Jaffe, D.B., Christodoulou, J., Fletcher, J.M., Bruno, D.L., Goldblatt, J., DiMauro, S., Thorburn, D.R., and Mootha, V.K., 2012. Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. *Science Translational Medicine*, **4**(118), 118ra10.
- Calvo, S.E., Clauser, K.R., and Mootha, V.K., 2015. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Research*, **44**(D1), pp.D1252-D1257.
- Camacho, J.A., Rioseco-Camacho, N., Andrade, D., Porter, J., and Kong, J., 2003. Cloning and characterization of human *ORNT2*: a second mitochondrial ornithine transporter that can rescue a defective *ORNT1* in patients with the hyperornithinemia-hyperammonemia-homocitrullinuria syndrome, a urea cycle disorder. *Molecular Genetics and Metabolism*, **79**(4), pp.257-271.
- Camacho, J.A., Obie, C., Biery, B., Goodman, B.K., Hu, C-A., Almashanu, S., Steel, G., Casey, R., Lambert, M., Mitchell, G.A., and Valle, D., 1999. Hyperornithinaemia-hyperammonaemia-homocitrullinuria syndrome is caused by mutations in a gene encoding a mitochondrial ornithine transporter. *Nature Genetics*, **22**(2), pp.151-158.
- Campos, R.K., Boratto, P.V., Assis, F.L., Aguiar, E.R.G.R., Silva, L.C.F., Albarnaz, J.D., Dornas, F.P., Trindade, G.S., Ferreira, P.P., Marques, J.T., Robert, C., Raoult, D., la Scola, B., and Abrahão, J.S., 2014. Samba virus: a novel mimivirus from a giant rain forest, the Brazilian Amazon. *Virology Journal*, **11**, p.95.
- Capaldi, R.A., 1990. Structure and function of cytochrome *c* oxidase. *Annual Review of Biochemistry*, **59**(1), pp.569-596.

- Cavalier-Smith, T., 1989. Molecular phylogeny. Archaeobacteria and Archezoa. *Nature*, **339**(6220), pp.100-101.
- Cermakian, N., Ikeda, T.M., Cedergren, R., and Gray, M.W., 1996. Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage. *Nucleic Acids Research*, **24**(4), pp.648-654.
- Chambers, J.W., Maguire, T.G., and Alwine, J.C., 2010. Glutamine metabolism is essential for human cytomegalovirus infection. *Journal of Virology*, **84**(4), pp.1867-1873.
- Chan, D.C., 2006. Mitochondrial fusion and fission in mammals. *Annual Review of Cell and Developmental Biology*, **22**(1), pp.79–99.
- Chang, E.S., Neuhof, M., Rubinstein, N.D., Diamant, A., Philippe, H., Huchon, D., and Cartwright, P., 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(48), pp.14912-14917.
- Chang, Y-F., 1976. Pipecolic acid pathway: The major lysine metabolic route in the rat brain. *Biochemical and Biophysical Research Communications*, **69**(1), pp.174-180.
- Chatzispyrou, I.A., Alders, M., Guerrero-Castillo, S., Perez, R.Z., Haagmans, M.A., Mouchiroud, L., Koster, J., Ofman, R., Bass, F., Waterham, H.R., Spelbrink, J.N., Auwerz, J., Mannens, M.M., Houtkooper, R.H., and Plomp, A.S., 2017. A homozygous missense mutation in *ERAL1*, encoding a mitochondrial rRNA chaperone, causes Perrault syndrome. *Human Molecular Genetics*, **26**(13), pp.2451-2550.
- Chen, F., Mackey, A.J., Vermunt, J.K., and Roos, D.S., 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One*, **2**(4), e383.
- Choudhary, C., Weinert, B.T., Nishida, Y., Verdin, E., and Mann, M., 2014. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nature Reviews: Molecular Cell Biology*, **15**(8), pp.536-550.
- Claros, M.G. and Vincens, P., 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. *European Journal of Biochemistry*, **241**(3), pp.779-786.

- Claus, C. and Liebert, U.G., 2014. A renewed focus on the interplay between viruses and mitochondrial metabolism. *Archives of Virology*, **159**(6), pp.1267-1277.
- Claverie, J-M., Grzela, R., Lartigue, A., Bernadac, A., Nitsche, S., Vacelet, J., Ogata, H., and Abergel, C., *et al.*, 2009. Mimivirus and Mimiviridae: giant viruses with an increasing number of potential hosts, including corals and sponges. *Journal of Invertebrate Pathology*, **101**(3), pp.172-180.
- Colson, P., de Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D.K., Cheng, X-W., Federici, B.A., van Etten, J.L., Koonin, E.V., la Scola, B., and Raoult, D., 2013. "Megavirales," a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Archives of Virology*, **158**(12), pp.2517-2521.
- Conant, G.C., and Wolfe, K.H., 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, **9**(12), pp.938-950.
- Contreras, L., Drago, I., Zampese, E., and Pozzan, T., 2010. Mitochondria: the calcium connection. *Biochimica et Biophysica Acta*, **1797**(6-7), pp.607-618.
- Corcoran, J.A., Saffran, H.A., Duguay, B.A., and Smiley, J.R, 2009. Herpes simplex virus UL12.5 targets mitochondria through a mitochondrial localization sequence proximal to the N terminus. *Journal of Virology*, **83**(6), pp.2601-2610.
- Csordás, G., Renken, C., Várnai, P., Walter, D., Buttle, K.F., Balla, T., Mannella, C.A., and Hajnóczky, G., 2006. Structural and functional features and significance of the physical linkage between ER and mitochondria. *The Journal of Cell Biology*, **174**(7), pp.915-921.
- Cuconati, A. and White, E., 2002. Viral homologs of BCL-2: role of apoptosis in the regulation of virus infection. *Genes & Development*, **16**(19), pp.2465-2478.
- Dalquen, D.A. and Dessimoz, C., 2013. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biology and Evolution*, **5**(10), pp.1800-1806.
- Danovaro, R., Dell'Anno, A., Pusceddu, A., Gambi, C., Heiner, I., and Kristensen, R.M., 2010. The first metazoa living in permanently anoxic conditions. *BMC Biology*, **8**, p.30.

- Darin, N., Oldfors, A., Moslemi, A-R., Holme, E., and Tulinius, M., 2001. The incidence of mitochondrial encephalomyopathies in childhood: clinical features and morphological, biochemical, and DNA abnormalities. *Annals of Neurology*, **49**(3), pp.377-383.
- De Magalhães, J.P. and Costa, J., 2009. A database of vertebrate longevity records and their relation to other life-history traits. *Journal of Evolutionary Biology*, **22**(8), pp.1770-1774.
- Dickerson, J.E. and Robertson, D.L., 2012. On the origins of Mendelian disease genes in man: the impact of gene duplication. *Molecular Biology and Evolution*, **29**(1), pp.61-69.
- DiMauro, S. and Schon, E.A., 2003. Mitochondrial respiratory-chain diseases. *The New England Journal of Medicine*, **348**(26), pp.2656-2668.
- Dolce, V., Scarcia, P., Iacopetta, D., and Palmieri, F., 2005. A fourth ADP/ATP carrier isoform in man: identification, bacterial expression, functional characterization and tissue distribution. *FEBS Letters*, **579**(3), pp.633-637.
- Dolezal, P., Dancis, A., Lesuisse, E., Sutak, R., Hrdý, I., Embley, T.M., and Tachezy, J., 2007. Frataxin, a conserved mitochondrial protein, in the hydrogenosome of *Trichomonas vaginalis*. *Eukaryotic Cell*, **6**(8), pp.1431-1438.
- Domazet-Lošo, T. and Tautz, D., 2008. An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular Biology and Evolution*, **25**(12), pp.2699-2707.
- Eddy, S.R., 1998. Profile hidden Markov models. *Bioinformatics*, **14**(9), pp.755-763.
- Edgar, R.C., 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, p.113.
- Edgar, R.C., 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), pp.1792-1797.
- Edvardson, S., Elbaz-Alon, Y., Jalas, C., Matlock, A., Patel, K., Labbé, K., Shaag, A., Jackman, J.E., and Elpeleg, O., 2016. A mutation in the *THGIL* gene in a family with cerebellar ataxia and developmental delay. *Neurogenetics*, **17**(4), pp.219-225.

- Ehrlich, R.S. and Colman, R.F., 1983. Separation, recombination, and characterization of dissimilar subunits of the DPN-dependent isocitrate dehydrogenase from pig heart. *The Journal of Biological Chemistry*, **258**(11), pp.7079-7086.
- Elstner, M., Andreoli, C., Ahting, U., Tetko, I., Klopstock, T., Meitinger, T., and Prokisch, H., 2009. MitoP2: an integrative tool for the analysis of the mitochondrial proteome. *Molecular Biotechnology*, **40**(3), pp.306-315.
- Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G., 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, **300**(4), pp.1005-1016.
- Eme, L., Sharpe, S.C., Brown, M.W., and Roger, A.J., 2014. On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harbor Perspectives in Biology*, **6**(8), a016139.
- Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A., 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**(6757), pp.86-90.
- Estévez, M., Skarda, J., Spencer, J., Banaszak, L., and Weaver, T.M., 2002. X-ray crystallographic and kinetic correlation of a clinically observed human fumarase mutation. *Protein Science*, **11**(6), pp.1552-1557.
- Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jazzal, B., Korninger, F., May, B., Milacic, M., Roca, C.D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wi, G., Stein, L., Hermjakob, H., and D'Eustachio, P., 2018. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, **46**(D1), pp.D649-D655.
- Farhan, S.M.K., Wang, J., Robinson, J.F., Lahiry, P., Siu, V.M., Prasad, C., Kronick, J.B., Ramsay, D.A., Rupa, C.A., and Hegele, R.A., 2014. Exome sequencing identifies *NFS1* deficiency in a novel Fe-S cluster disease, infantile mitochondrial complex II/III deficiency. *Molecular Genetics & Genomic Medicine*, **2**(1), pp.73-80.
- Fernandez-Moran, H., 1962. Cell-membrane ultrastructure. Low-temperature electron microscopy and x-ray diffraction studies of lipoprotein components in lamellar systems. *Circulation*, **26**, pp.1039-1065.

- Fernández-Murray, J.P., Prykhozhiy, S.V., Dufay, J.N., Steele, S.L., Gaston, D., Nasrallah, G.K., Coombs, A.J., Liwski, R.S., Fernandez, C.V., Berman, J.N., and McMaster, C.R., 2016. Glycine and folate ameliorate models of congenital sideroblastic anemia. *PLoS Genetics*, **12**(1), e1005783.
- Fiermonte, G., Dolce, V., Palmieri, L., Ventura, M., Runswick, M.J., Palmieri, F., and Walker, J.E., 2001. Identification of the human mitochondrial oxodicarboxylate carrier. Bacterial expression, reconstitution, functional characterization, tissue distribution, and chromosomal location. *The Journal of Biological Chemistry*, **276**(11), pp.8225-8230.
- Fiermonte, G., Palmieri, L., Dolce, V., Lasorsa, F.M., Palmieri, F., Runswick, M.J., and Walker, J.E., 1998. The sequence, bacterial expression, and functional reconstitution of the rat mitochondrial dicarboxylate transporter cloned via distant homologs in yeast and *Caenorhabditis elegans*. *The Journal of Biological Chemistry*, **273**(38), pp.24754-24759.
- Fitch, W.M., 2000. Homology: a personal view on some of the problems. *Trends in Genetics*, **16**(5), pp.227-231.
- Fitzpatrick, D.A., Creevey, C.J., and McInerney, J.O., 2006. Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales. *Molecular Biology and Evolution*, **23**(1), pp.74-85.
- Floyd, B.J., et al., 2016. Mitochondrial protein interaction mapping identifies regulators of respiratory chain function. *Molecular Cell*, **63**(4), pp.621-632.
- Formosa, L.E., Mimaki, M., Frazier, A.E., McKenzie, M., Stait, T.L., Thorburn, D.R., Stroud, D.A., and Ryan, M.T., 2015. Characterization of mitochondrial *FOXRED1* in the assembly of respiratory chain complex I. *Human Molecular Genetics*, **24**(10), pp.2952-2965.
- French, J.B., Jones, S.A., Deng, H., Pedley, A.M., Kim, D., Chan, C.Y., Hu, H., Pugh, R.J., Zhao, H., Zhang, Y., Huang, T.J., Fang, Y., Zhuang, X., and Benkovic, S.J., 2016. Spatial colocalization and functional link of purinosomes with mitochondria. *Science*, **351**(6274), pp.733-737.
- Froese, D.S. and Gravel, R.A., 2010. Genetic disorders of vitamin B12 metabolism: eight complementation groups – eight genes. *Expert Reviews in Molecular Medicine*, **12**, e37.

- Fukasawa, Y., Tsuji, J., Fu, S-C., Tomii, K., Horton, P., and Imai, K., 2015. MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Molecular & Cellular Proteomics*, **14**(4), pp.1113-1126.
- Fukumura, S., Ohba, C., Watanabe, T., Minagawa, K., Shimura, M., Murayama, K., Ohtakem A., Saitsu, H., Matsumoto, N., and Tsutsumi, H., 2015. Compound heterozygous *GFM2* mutations with Leigh syndrome complicated by arthrogryposis multiplex congenita. *Journal of Human Genetics*, **60**(9), pp.509-513.
- Furney, S.J., Albà, M.M., and López-Bigas, N., 2006. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. *BMC Genomics*, **7**, p.165.
- Gabaldón, T. and Huynen, M.A., 2007. From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. *PLoS Computational Biology*, **3**(11), e219.
- Gabrielson, M., Reizer, E., Stål, O., and Tina, E., 2016. Mitochondrial regulation of cell cycle progression through *SLC25A43*. *Biochemical and Biophysical Research Communications*, **469**(4), pp.1090-1096.
- Galluzzi, L., Brenner, C., Morselli, E., Touat, Z., and Kroemer, G., 2008. Viral control of mitochondrial apoptosis. *PLoS Pathogens*, **4**(5), e1000018.
- Giacomello, M. and Pellegrini, L., 2016. The coming of age of the mitochondria-ER contact: a matter of thickness. *Cell Death and Differentiation*, **23**(9), pp.1417-1427.
- Giaever, G. *et al.*, 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**(6896), pp.387-391.
- Gilbert, C., Peccoud, J., Chateigner, A., Moumen, B., Cordaux, R., and Herniou, E.A., 2016. Continuous influx of genetic material from host to virus populations. *PLoS Genetics*, **12**(2), e1005838.
- Goldberg, A.V., Molik, S., Tsaousis, A.D., Neumann, K., Kuhnke, G., Delbac, F., Vivares, C.P., Hirt, R.P., Lill, R., and Embley, T.M., 2008. Localization and functionality of microsporidian iron-sulphur cluster assembly proteins. *Nature*, **452**(7187), pp.624-628.

- Gorman, G.S., Schaefer, A.M., Ng, Y., Gomez, G., Blakely, E.L., Alston, C.L., Feeney, C., Horvath, R., Yu-Wai-Man, P., Chinnery, P.F., Taylor, R.W., Turnbull, D.M., and McFarland, R., 2015. Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Annals of Neurology*, **77**(5), pp.753-759.
- Gray, K.A., Seal, R.L., Tweedie, S., Wright, M.W., and Bruford, E.A., 2016. A review of the new HGNC gene family resource. *Human Genomics*, **10**, p.6.
- Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., and Bruford, E.A., 2014. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Research*, **43**(D1), pp.D1079-D1085.
- Gray, M.W., 2015. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(33), pp.10133-10138.
- Gribaldo, S., Poole, A.M., Daubin, V., Forterre, P., and Brochier-Armanet, C., 2010. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nature Reviews: Microbiology*, **8**(10), pp.743-752.
- Gross, A., McDonnell, J.M., and Korsmeyer, S.J., 1999. BCL-2 family members and the mitochondria in apoptosis. *Genes & Development*, **13**(15), pp.1899-1911.
- Guda, P., Subramaniam, S., and Guda, C., 2007. Mitoproteome: human heart mitochondrial protein sequence database. *Methods in Molecular Biology*, **357**, pp.375-383.
- Guernsey, D.L., *et al.*, 2009. Mutations in mitochondrial carrier family gene *SLC25A38* cause nonsyndromic autosomal recessive congenital sideroblastic anemia. *Nature Genetics*, **41**(6), pp.651-653.
- Guerrero-Castillo, S., Baertling, F., Kownatzki, D., Wessels, H.J., Arnold, S., Brandt, U., and Nijtmans, L., 2017. The assembly pathway of mitochondrial respiratory chain complex I. *Cell Metabolism*, **25**(1), pp.128-139.
- Guindon, S., Dufayard, J-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**(3), pp.307-321.

- Gutiérrez-Aguilar, M. and Baines, C.P., 2013. Physiological and pathological roles of mitochondrial SLC25 carriers. *The Biochemical Journal*, **454**(3), pp.371-386.
- Haguenauer, A., Raimbault, S., Masscheleyn, S., Gonzalez-Barroso, M.M., Criscuolo, F., Plamondon, J., Miroux, B., Ricquier, D., Richard, D., Bouillard, F., and Pecqueur, C., 2005. A new renal mitochondrial carrier, *KMCP1*, is up-regulated during tubular cell regeneration and induction of antioxidant enzymes. *The Journal of Biological Chemistry*, **280**(23), pp.22036-22043.
- Haitina, T., Linblom, J., Renström, T., and Fredriksson, R., 2006. Fourteen novel human members of mitochondrial solute carrier family 25 (SLC25) widely expressed in the central nervous system. *Genomics*, **88**(6), pp.779-790.
- Hansford, R.G. and Johnson, R.N., 1975. The steady state concentrations of coenzyme A-SH and coenzyme A thioester, citrate, and isocitrate during tricarboxylate cycle oxidations in rabbit heart mitochondria. *The Journal of Biological Chemistry*, **250**(21), pp.8361-8375.
- Henriksen, P., Wagner, S.A., Weinert, B.T., Sharma, S., Bačinskaja, G., Rehman, M., Juffer, A.H., Walther, T.C., Lisby, M., and Choudhary, C., 2012. Proteome-wide analysis of lysine acetylation suggests its broad regulatory scope in *Saccharomyces cerevisiae*. *Molecular & Cellular Proteomics*, **11**(11), pp.1510-1522.
- Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Spooner, S.B.W., Kulesha, E., Yates, A., and Flicek, P., 2016. Ensembl comparative genomics resources. *Database*, **2016**, baw053.
- Higashino, K., Fujioka, M., Aoki, T., and Tamamura, Y., 1967. Metabolism of lysine in rat liver. *Biochemical and Biophysical Research Communications*, **29**(1), pp.95-100.
- Higashino, K., Tsukada, K., and Lieberman, I., 1965. Saccharopine, a product of lysine breakdown by mammalian liver. *Biochemical and Biophysical Research Communications*, **20**(3), pp.285-290.
- Houten, S.M., Violante, S., Ventura, F.V., and Wanders, R.J.A., 2016. The biochemistry and physiology of mitochondrial fatty acid β -oxidation and its genetic disorders. *Annual Review of Physiology*, **78**(1), pp.23-44.

- Howe, K.L., Bolt, B.J., Shafie, M., Kersey, P., and Berriman, M., 2017. WormBase ParaSite – a comprehensive resource for helminth genomics. *Molecular and Biochemical Parasitology*, **215**, pp.2-10.
- Huang, D.W., Sherman, B.T., and Lempicki, R.A., 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**(1), pp.1-13.
- Huang, D.W., Sherman, B.T., and Lempicki, R.A., 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**(1), pp.44-57.
- Hulsen, T., Huynen, M.A., de Vlieg, J., and Groenen, P.M.A., 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*, **7**(4), p.R31.
- Hung, V., Zou, P., Rhee, H-W., Udeshi, N.D., Cracan, V., Svinkina, T., Carr, S.A., Mootha, V.K., and Ting, A.Y., 2014. Proteomic mapping of the human mitochondrial intermembrane space in live cells via ratiometric APEX tagging. *Molecular Cell*, **55**(2), pp.332-341.
- Huss, J.M. and Kelly, D.P., 2005. Mitochondrial energy metabolism in heart failure: a question of balance. *The Journal of Clinical Investigation*, **115**(3), pp.547-555.
- Huttlin, E.L. *et al.*, 2017. Architecture of the human interactome defines protein communities and disease networks. *Nature*, **545**(7655), pp.505-509.
- Huttlin, E.L. *et al.*, 2015. The BioPlex network: a systematic exploration of the human interactome. *Cell*, **162**(2), pp.425-440.
- Ikeda, F., *et al.*, 2017. Exome sequencing identified *RPS15A* as a novel causative gene for Diamond-Blackfan anemia. *Haematologica*, **102**(3), e93-e96.
- Islamaj Doğan, R., Kim, S., Chatr-aryamontri, A., Chang, C.S., Oughtred, R., Rust, J., Wilbur, W.J., Comeau, D.C., Dolinski, K., and Tyers, M., 2017. The BioC-BioGRID corpus: full text articles annotated for curation of protein–protein and genetic interactions. *Database*, **2017**, baw147.

- James, A.M., Hoogewijs, K., Logan, A., Hall, A.R., Ding, S., Fearnley, I.M., and Murphy, M.P., 2017. Non-enzymatic *N*-acetylation of lysine residues by acetyl-CoA often occurs via a proximal *S*-acetylated thiol intermediate sensitive to glyoxalase II. *Cell Reports*, **18**(9), pp.2105-2112.
- Jimenez-Sanchez, G., Childs, B., and Valle, D., 2001. Human disease genes. *Nature*, **409**(6822), pp.853-855.
- Jothi, R., Przytycka, T.M., and Aravind, L., 2007. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics*, **8**, p.173.
- Kanaani, J., Cianciaruso, C., Phelps, E.A., Pasquier, M., Brioude, E., Billestrup, N., and Baekkeskov, S., 2015. Compartmentalization of GABA synthesis by *GAD67* differs between pancreatic beta cells and neurons. *PloS One*, **10**(2), e0117130.
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, **44**(D1), pp.D457-D462.
- Kanfi, Y., Naiman, S., Amir, G., Peshti, V., Zinman, G., Nahum, L., Bar-Joseph, Z., and Cohen, H.Y., 2012. The sirtuin *SIRT6* regulates lifespan in male mice. *Nature*, **483**(7388), pp.218-221.
- Kariithi, H.M., Yao, X., Yu, F., Teal, P.E., Verhoeven, C.P., and Boucias, D.G., 2017. Responses of the housefly, *Musca domestica*, to the hytrosavirus replication: impacts on host's vitellogenesis and immunity. *Frontiers in Microbiology*, **8**, p.583.
- Karlberg, O., Canbäck, B., Kurland, C.G., and Andersson, S.G.E., 2000. The dual origin of the yeast mitochondrial proteome. *Yeast*, **17**(3), pp.170-187.
- Karniely, S., Weekes, M.P., Antrobus, R., Rorbach, J., van Haute, L., Umraniya, Y., Smith, D.L., Stanton, R.J., Minczuk, M., Lehner, P.J., and Sinclair, J.H., 2016. Human cytomegalovirus infection upregulates the mitochondrial transcription and translation machineries. *mBio*, **7**(2), e00029.

- Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S.C., Petrželková, R., Eme, L., Novák, L., Žárský, V., Barlow, L.D., Herman, E.K., Soukal, P., Hroudová, M., Doležal, P., Stairs, C.W., Roger, A.J., Eliáš, M., Dacks, J.B., Vlček, Č., and Hampl, V., 2016. A eukaryote without a mitochondrial organelle. *Current Biology*, **26**(10), pp.R410-R412.
- Kashiwaya, Y., Pawlosky, R., Markis, W., King, M.T., Bergman, C., Srivastava, S., Murray, A., Clarke, K., and Veech, R.L., 2010. A ketone ester diet increases brain malonyl-CoA and Uncoupling proteins 4 and 5 while decreasing food intake in the normal Wistar Rat. *The Journal of Biological Chemistry*, **285**(34), pp.25950-25956.
- Keeling, P.J. and Palmer, J.D., 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, **9**(8), pp.605-618.
- Keeling, P.J. and Slamovits, C.H., 2005. Causes and effects of nuclear genome reduction. *Current Opinion in Genetics & Development*, **15**(6), pp.601-608.
- Keen, B.A., Bailey, L.J., Jozwiakowski, S.K., and Doherty, A.J., 2014. Human PrimPol mutation associated with high myopia has a DNA replication defect. *Nucleic Acids Research*, **42**(19), pp.12102-12111.
- Kernohan, K.D., *et al.*, 2017. Matchmaking facilitates the diagnosis of an autosomal-recessive mitochondrial disease caused by biallelic mutation of the tRNA isopentenyltransferase (*TRIT1*) gene. *Human Mutations*, **38**(5), pp.511-516.
- Kikuchi, G., Motokawa, Y., Yoshida, T., and Hiraga, K., 2008. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, **84**(7), pp.246-263.
- Kim, D., Nguyen, M.D., Dobbin, M.M., Fischer, A., Sananbenesi, F., Rodgers, J.T., Delalle, I., Baur, J.A., Sui, G., Armour, S.M., Puigserver, P., Sinclair, D.A., and Tsai, L-H., 2007. *SIRT1* deacetylase protects against neurodegeneration in models for Alzheimer's disease and amyotrophic lateral sclerosis. *The EMBO Journal*, **26**(13), pp.3169-3179.
- Kim, H.J., Khalimonchuk, O., Smith, P.M., and Winge, D.R., 2012. Structure, function, and assembly of heme centers in mitochondrial respiratory complexes. *Biochimica et Biophysica Acta*, **1823**(9), pp.1604-1616.

- King, M.S., Kerr., M., Crichton, P.G., Springett, R., and Kunji E.R.S., 2016. Formation of a cytoplasmic salt bridge network in the matrix state is a fundamental step in the transport mechanism of the mitochondrial ADP/ATP carrier. *Biochimica et Biophysica Acta*, **1857**(1), pp.14-22.
- Kishita, Y., *et al.*, 2015. Intra-mitochondrial methylation deficiency due to mutations in *SLC25A26*. *American Journal of Human Genetics*, **97**(5), pp.761-768.
- Kondrashov, F.A. and Koonin, E.V., 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends in Genetics*, **20**(7), pp.287-290.
- Koonin, E.V., 2005a. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, **39**, pp.309-338.
- Koonin, E.V., 2005b. Virology: Gulliver among the Lilliputians. *Current Biology*, **15**(5), pp.R167-R169.
- Krishnamurthy, P.C., Du, G., Fukuda, Y., Sun, D., Sampath, J., Mercer, K.E., Wang, J., Sosa-Pineda, B., Murti, K.G., and Schuetz, J.D., 2006. Identification of a mammalian mitochondrial porphyrin transporter. *Nature*, **443**(7111), pp.586-589.
- Kristensen, D.M., Cai, X., and Mushegian, A., 2011. Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *Journal of Bacteriology*, **193**(8), pp.1806-1814.
- Krumm, N., Sudmant, P.H., Ko, A., O’Roak, B.J., Malig, M., Coe, B.P., NHLBI Exome Sequencing Project, Quinlan, A.R., Nickerson, D.A., and Eichler, E.E., 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Research*, **22**(8), pp.1525-1532.
- Kruse, I., Maclean, A.E., Hill, L., and Balk, J., 2018. Genetic dissection of cyclic pyranopterin monophosphate biosynthesis in plant mitochondria. *The Biochemical Journal*, **475**(2), pp.495-509.

- Kuczyńska-Wiśnik, D. I., Moruno-Algara, M., Stojowska-Swędryńska, K., and Laskowska, E., 2016. The effect of protein acetylation on the formation and processing of inclusion bodies and endogenous protein aggregates in *Escherichia coli* cells. *Microbial Cell Factories*, **15**(1), p.189.
- Kunji, E.R.S., 2004. The role and structure of mitochondrial carriers. *FEBS Letters*, **564**(3), pp.239-244.
- Kunji, E.R.S., Aleksandrova, A., King, M.S., Majd, H., Ashton, V.L., Cerson, E., Springett, R., Kibalchenko, M., Tavoulari, S., Crichton, P.G., and Ruprecht, J.J., 2016. The transport mechanism of the mitochondrial ADP/ATP carrier. *Biochimica et Biophysica Acta*, **1863**(10), pp.2379-2393.
- Kurland, C.G. and Andersson, S.G., 2000. Origin and evolution of the mitochondrial proteome. *Microbiology and Molecular Biology Reviews*, **64**(4), pp.786-820.
- Laffel, L., 1999. Ketone bodies: a review of physiology, pathophysiology and application of monitoring to diabetes. *Diabetes Metabolism Research and Reviews*, **15**(6), pp.412-426.
- Lane, N. and Martin, W., 2010. The energetics of genome complexity. *Nature*, **467**(7318), pp.929-934.
- Lee, A.J., Cai, M.X., Thomas, P.E., Conney, A.H., and Zhu, B.T., 2003. Characterization of the oxidative metabolites of 17- β -estradiol and estrone formed by 15 selectively expressed human cytochrome p450 isoforms. *Endocrinology*, **144**(8), pp.3382-3398.
- Lee, C.F. and Tian, R., 2015. Mitochondrion as a target for heart failure therapy – role of protein lysine acetylation. *Circulation Journal*, **79**(9), pp.1863-1870.
- Lek, M. *et al.*, 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**(7616), pp.285-291.
- Letunic, I. and Bork, P., 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, **44**(W1), pp.W242-W245.
- Li, J. and Zhang, Q., 2015. *PRIMPOL* mutation: functional study does not always reveal the truth. *Investigative Ophthalmology & Visual Science*, **56**(2), pp.1181-1182.

- Li, L., Stoeckert, C.J., and Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**(9), pp.2178-2189.
- Liao, B-Y. and Zhang, J., 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(19), pp.6987-6992.
- Lin, S.J., Defossez, P.A., and Guarente, L., 2000. Requirement of NAD and *SIR2* for life-span extension by calorie restriction in *Saccharomyces cerevisiae*. *Science*, **289**(5487), pp.2126-2128.
- Lionel, A.C. *et al.*, 2018. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genetics in Medicine*, **20**(4), pp.435-443.
- Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., Ghabrial, S.A., Li, G., Yi, X., and Jiang, D., 2011. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evolutionary Biology*, **11**, p.276.
- Liu, H., Fu, Y., Jiang, D., Li, G., Xie, J., Cheng, J., Peng, Y., Ghabrial, S.A., and Yi, X., 2010. Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. *Journal of Virology*, **84**(22), pp.11876-11887.
- Llopis, J., McCaffery, M., Miyawaki, A., Farquhar, M.G., and Tsien, R.Y., 1998. Measurement of cytosolic, mitochondrial, and Golgi pH in single living cells with green fluorescent proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(12), pp.6803-6808.
- Loganathanaraj, R. and Atwi, M., 2007. Towards validating the hypothesis of phylogenetic profiling. *BMC Bioinformatics*, **8**(S7), p.S25.
- Lombard, D.B., Alt, F.W., Cheng, H-L., Bunkenborg, J., Streeper, R.S., Mostoslavsky, R., Kim, J., Tancopulous, G., Valenzuela, D., Murphy, A., Yang, Y., Chen, Y., Hirschey, M.D., Bronson, R.T., Haigis, M., Guarente, L.P., Farese Jr., R.V., Weissman, S., Verdin, E., and Schwer, B., 2007. Mammalian *Sir2* homolog *SIRT3* regulates global mitochondrial lysine acetylation. *Molecular and Cellular Biology*, **27**(24), pp.8807-8814.

- López-Bigas, N. and Ouzounis, C.A., 2004. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Research*, **32**(10), pp.3108-3114.
- Lyons, A.M., Ardisson, A., Reyes, A., Robinson, A.J., Moroni, I., Ghezzi, D., Fernandez-Vizarra, E., and Zeviani, M., 2016. *COA7 (C1orf163/RESA1)* mutations associated with mitochondrial leukoencephalopathy and cytochrome c oxidase deficiency. *Journal of Medical Genetics*, **53**(12), pp.846-849.
- Mao, G., Tan, J., Gao, W., Shi, Y., Cui, M-Z., and Xu, X., 2008. Both the I-terminal fragment and the protein-protein interaction domain (PDZ domain) are required for the pro-apoptotic activity of presenilin-associated protein PSAP. *Biochimica et Biophysica Acta*, **1780**(4), pp.696-708.
- Marchler-Bauer, A., Bo, Y., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Geer, L.Y., and Bryant, S.H. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, **45**(D1), pp.D200-D203.
- Marcotte, E.M., Xenarios, I., van der Bliek, A.M., and Eisenberg, D., 2000. Localizing proteins in the cell from their phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(22), pp.12115-12120.
- Margolin, W., 2005. FtsZ and the division of prokaryotic cells and organelles. *Nature Reviews: Molecular Cell Biology*, **6**(11), pp.862-871.
- Marino, S.M. and Gladyshev, V.N., 2010. Cysteine function governs its conservation and degeneration and restricts its utilization on protein surfaces. *Journal of Molecular Biology*, **404**(5), pp.902-916.
- Masters, B.S., Stohl, L.L., and Clayton, D.A., 1987. Yeast mitochondrial RNA polymerase is homologous to those encoded by bacteriophages T3 and T7. *Cell*, **51**(1), pp.89-99.
- Matthews, P.M., Brown, R.M., Otero, L.J., Marchington, D.R., LeGris, M., Howes, R., Meadows, L.S., Shevell, M., Scriver, C.R., and Brown, G.K., 1994. Pyruvate dehydrogenase deficiency. Clinical presentation and molecular genetic characterization of five new patients. *Brain*, **117**(3), pp.435-443.

- Maxwell, E.K., Schnitzler, C.E., Havlak, P., Putnam, N.H., Nguyen, A-D., Moreland, R.T., and Baxeavanis, A.D., 2014. Evolutionary profiling reveals the heterogeneous origins of classes of human disease genes: implications for modeling disease genetics in animals. *BMC Evolutionary Biology*, **14**, p.212.
- Małecki, J., Ho, A.Y.Y., Moen, A., Dahl, H-A., and Falnes, P.Ø., 2015. Human *METTL20* is a mitochondrial lysine methyltransferase that targets the β subunit of electron transfer flavoprotein (ETF β) and modulates its activity. *The Journal of Biological Chemistry*, **290**(1), pp.423-434.
- McBride, H.M., Neuspiel, M., and Wasiak, S., 2006. Mitochondria: more than just a powerhouse. *Current Biology*, **16**(14), pp.R551-R560.
- McCarthy, E.A., Titus, S.A., Taylor, S.M., Jackson-Cook, C., and Moran, R.G., 2004. A mutation inactivating the mitochondrial inner membrane folate transporter creates a glycine requirement for survival of Chinese hamster cells. *The Journal of Biological Chemistry*, **279**(32), pp.33829-33836.
- McCrae, J.F., *et al.*, 2017. Prevalence and architecture of *de novo* mutations in developmental disorders. *Nature*, **542**(7642), pp.433-438.
- McGarry, J.D., Guest, M.J., and Foster, D.W., 1970. Ketone body metabolism in the ketosis of starvation and alloxan diabetes. *The Journal of Biological Chemistry*, **245**(17), pp.4382-4390.
- Mckenzie, M. and Ryan, M.T., 2010. Assembly factors of human mitochondrial complex I and their defects in disease. *IUBMB Life*, **62**(7), pp.497-502.
- Meisinger, C., Sickmann, A., and Pfanner, N., 2008. The mitochondrial proteome: from inventory to function. *Cell*, **134**(1), pp.22-24.
- Mendel, R.R., 2013. The molybdenum cofactor. *The Journal of Biological Chemistry*, **288**(19), pp.13165-13172.
- Menezes, M.J., *et al.*, 2015. Mutation in mitochondrial ribosomal protein S7 (*MRPS7*) causes congenital sensorineural deafness, progressive hepatic and renal failure, and lactic acidemia. *Human Molecular Genetics*, **24**(8), pp.2297-2307.

- Mentel, M., Tielens, A.G.M., and Martin, W.F., 2016. Animals, anoxic environments, and reasons to go deep. *BMC Biology*, **14**, p.44.
- Mezl, V.A. and Knox, W.E., 1976. Properties and analysis of a stable derivative of pyrroline-5-carboxylic acid for use in metabolic studies. *Analytical Biochemistry*, **74**(2), pp.430-440.
- Mi-Ichi, F., Yousuf, M.A., Nakada-Tsukui, K., and Nozaki, T., 2009. Mitosomes in *Entamoeba histolytica* contain a sulfate activation pathway. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(51), pp.21731-21736.
- Michel, V. and Bakovic, M., 2009. The solute carrier 44A1 is a mitochondrial protein and mediates choline transport. *FASEB Journal*, **23**(8), pp.2749-2758.
- Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., and Ogata, H., 2016. Linking virus genomes with host taxonomy. *Viruses*, **8**(3), p.66.
- Miller, W.L., 2013. Steroid hormone synthesis in mitochondria. *Molecular and Cellular Endocrinology*, **379**(1-2), pp.62-73.
- Min, S-W, Chen, X., Tracy, T.E., Li, Y., Zhou, Y., Wang, C., Shirakawa, K., Minami, S.S., Defensor, E., Mok, S.A., Sohn, P.D., Schilling, B., Cong, X., Ellerby, L., Gibson, B.W., Johnson, J., Krogan, N., Shamloo, M., Gestwicki, J., Masliah, E., Verdin E., and Gan, L., 2015. Critical role of acetylation in tau-mediated neurodegeneration and cognitive deficits. *Nature Medicine*, **21**(10), pp.1154-1162.
- Monier, A., Pagarate, A., de Vargas, C., Allen, M.J., Read, B., Claverie, J-M., and Ogata, H., 2009. Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Research*, **19**(8), pp.1441-1449.
- Monier, A., Claverie, J-M., and Ogata, H., 2007. Horizontal gene transfer and nucleotide compositional anomaly in large DNA viruses. *BMC Genomics*, **8**, p.456.
- Monné, M., Robinson, A.J., Boes, C., Harbour, M.E., Fearnley, I.M., and Kunji, E.R.S., 2007. The mimivirus genome encodes a mitochondrial carrier that transports dATP and dTTP. *Journal of Virology*, **81**(7), pp.3181-3186.

- Moreau, Y. and Tranchevent, L.-C., 2012. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews: Genetics*, **13**(8), pp.523-536.
- Moreno-Hagelsieb, G. and Latimer, K., 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, **24**(3), pp.319-324.
- Morita, M., Prudent, J., Basu, K., Goyon, V., Katsumura, S., Hulea, L., Pearl, D., Siddiqui, N., Strack, S., McGuirk, S., St-Pierre, J., Larsson, O., Topisirovic, I., Vali, H., McBride, H.M., Bergeron, J.J., and Sonenberg, N., 2017. mTOR controls mitochondrial dynamics and cell survival via *MTFP1*. *Molecular Cell*, **67**(6), pp.922-935.
- Motenko, H., Neuhauser, S.B., O'Keefe, M., and Richardson, J.E., 2015. MouseMine: a new data warehouse for MGI. *Mammalian Genome*, **26**(7-8), pp.325-330.
- Munger, J., Bajad, S.U., Collier, H.A., Shenk, T., and Rabinowitz, J.D., 2006. Dynamics of the cellular metabolome during human cytomegalovirus infection. *PLoS Pathogens*, **2**(12), e132.
- Munger, J., Bennett, B.D., Parikh, A., Feng, X.-J., McArdle, J., Rabitz, H.A., Shenk, T., and Rabinowitz, J.D., 2008. Systems-level metabolic flux profiling identifies fatty acid synthesis as a target for antiviral therapy. *Nature Biotechnology*, **26**(10), pp.1179-1186.
- Musante, L., Püttmann, L., Kahrizi, K., Garshasbi, M., Hu, H., Stehr, H., Lipkowitz, B., Otto, S., Jensen, L.R., Tzschach, A., Jamali, P., Wienker, T., Najmabadi, H., Ropers, H.H., and Kuss, A.W., 2017. Mutations of the aminoacyl-tRNA-synthetases *SARS* and *WARS2* are implicated in the etiology of autosomal recessive intellectual disability. *Human Mutation*, **38**(6), pp.621-636.
- Müller, M., 1993. The hydrogenosome. *Journal of General Microbiology*, **139**(12), pp.2879-2889.
- Müller, M. and Martin, W., 1999. The genome of *Rickettsia prowazekii* and some thoughts on the origin of mitochondria and hydrogenosomes. *BioEssays*, **21**(5), pp.377-381.
- Narasimhan, V.M., *et al.*, 2016. Health and population effects of rare gene knockouts in adult humans with related parents. *Science*, **352**(6284), pp.474-477.

- Nazli, A., Safdar, A., Saleem, A., Akhtar, M., Brady, L.I., Schwartzentruber, J., and Tarnopolsky, M.A., 2017. A mutation in the *TMEM65* gene results in mitochondrial myopathy with severe neurological manifestations. *European Journal of Human Genetics*, **25**(6), pp.744-751.
- Nedergaard, J., Golozoubova, V., Matthias, A., Asadi, A., Jacobsson, A., and Cannon, B., 2001. *UCP1*: the only protein able to mediate adaptive non-shivering thermogenesis and metabolic inefficiency. *Biochimica et Biophysica Acta*, **1504**(1), pp.82-106.
- Nehrt, N.L., Clark, W.T., Radivojac, P., and Hahn, M.W., 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Computational Biology*, **7**(6), e1002073.
- Nelson, D.R., Felix, C.M., and Swanson, J.M., 1998. Highly conserved charge-pair networks in the mitochondrial carrier family. *Journal of Molecular Biology*, **277**(2), pp.285-308.
- Nunnari, J. and Suomalainen, A., 2012. Mitochondria: in sickness and in health. *Cell*, **148**(6), pp.1145-1159.
- O'Brien, T.W., 2002. Evolution of a protein-rich mitochondrial ribosome: implications for human genetic disease. *Gene*, **286**(1), pp.73-79.
- Ogilvie, I., Kennaway, N.G., and Shoubridge, E.A., 2005. A molecular chaperone for mitochondrial complex I assembly is mutated in a progressive encephalopathy. *The Journal of Clinical Investigation*, **115**(10), pp.2784-2792.
- Ohta, A. and Nishiyama, Y., 2011. Mitochondria and viruses. *Mitochondrion*, **11**(1), pp.1-12.
- Onyango, I.G., Khan, S.M., and Bennett Jr, J.P., 2017. Mitochondria in the pathophysiology of Alzheimer's and Parkinson's diseases. *Frontiers in Bioscience*, **22**(5), pp.854-872.
- Ostergaard, E., Weraarpachai, W., Ravn, K., Born, A.P., Jønson, L., Duno, M., Wibrand, F., Shoubridge, E.A., and Vissing, J., 2015. Mutations in *COA3* cause isolated complex IV deficiency associated with neuropathy, exercise intolerance, obesity, and short stature. *Journal of Medical Genetics*, **52**(3), pp.203-207.

- Owen, O.E., Morgan, A.P., Kemp, H.G., Sullivan, J.M., Herrera, M.G., and Cahill Jr., G.F., 1967. Brain metabolism during fasting. *The Journal of Clinical Investigation*, **46**(10), pp.1589-1595.
- Pagliarini, D.J., Calvo, S.E., Chang, B., Sheth, S.A., Vafai, S.B., Ong, S-E., Walford, G.A., Sugiana, C., Boneh, A., Chen, W.K., Hill, D.E., Vidal, M., Evans, J.G., Thorburn, D.R., Carr, S.A., and Mootha, V.K., 2008. A mitochondrial protein compendium elucidates complex I disease biology. *Cell*, **134**(1), pp.112-123.
- Palmieri, F., 2013. The mitochondrial transporter family SLC25: identification, properties and physiopathology. *Molecular Aspects of Medicine*, **34**(2-3), pp.465-484.
- Palmieri, L., Palmieri, F., Runswick, M.J., and Walker, J.E., 1996. Identification by bacterial expression and functional reconstitution of the yeast genomic sequence encoding the mitochondrial dicarboxylate carrier protein. *FEBS Letters*, **399**(3), pp.299-302.
- Palmieri, L., Agrimi, G., Runswick, M.J., Fearnley, I.M., Palmieri, F., and Walker, J.E., 2001a. Identification in *Saccharomyces cerevisiae* of two isoforms of a novel mitochondrial transporter for 2-oxoadipate and 2-oxoglutarate. *The Journal of Biological Chemistry*, **276**(3), pp.1916-1922.
- Palmieri, L., Pardo, B., Lasorsa, F.M., del Arco, A., Kobayashi, K., Iijima, M., Runswick, M.J., Walker, J.E., Saheki, T., Satrústegui, J., and Palmieri, F., 2001b. Citrin and aralar1 are Ca^{2+} -stimulated aspartate/glutamate transporters in mitochondria. *The EMBO Journal*, **20**(18), pp.5060-5069.
- Pebay-Peyroula, E., Dahout-Gonzalez, C., Kahn, R., Trézéguet, V., Lauquin, G.J-M., and Brandolin, G., 2003. Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. *Nature*, **426**(6962), pp.39-44.
- Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(8), pp.4285-4288.

- Perez, Y., Shorer, Z., Liani-Leibson, K., Chabosseu, P., Kadir, R., Volodarsky, M., Halperin, D., Barber-Zucker, S., Shalev, H., Schrieiber, R., Gradstein, L., Gurevich, E., Zarivach, R., Rutter, G.A., Landau, D., and Birk, O.S., 2017. *SLC30A9* mutation affecting intracellular zinc homeostasis causes a novel cerebro-renal syndrome. *Brain*, **140**(4), pp.928-939.
- Pires-daSilva, A. and Sommer, R.J., 2003. The evolution of signalling pathways in animal development. *Nature Reviews: Genetics*, **4**(1), pp.39-49.
- Pitceathly, R.D.S., Rahman, S., Wedatilake, Y., Polke, J.M., Cirak, S., Foley, A.R., Sailer, A., Hurles, M.E., Stalker, J., Hargreaves, I., Woodward, C.E., Sweeney, M.G., Muntoni, F., Houlden, H., UK10K Consortium, Taanmen, J-W., and Hanna, M.G., 2013. *NDUFA4* mutations underlie dysfunction of a cytochrome *c* oxidase subunit linked to human neurological disease. *Cell Reports*, **3**(6), pp.1795-1805.
- Podlutzky, A.J., Khritankov, A.M., Ovodov, N.D., and Austad, S.N., 2005. A new field record for bat longevity. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, **60**(11), pp.1366-1368.
- Poole, O.V., Fernandez-Vizarra, E., Turner, C., Clarke, B., Bugiardini, E., Barbosa, I.A., Deshpande, C., Hargreaves, I.P., Woodward, C.E., Sweeney, M.G., Poulton, J., Simpson, M.A., Houlden, H., Zeviani, M., Hanna, M.G., and Pitceathly, R.D.S., 2016. *TRIAP1* mutations are a cause of reversible infantile bulbar failure with subsequent progressive adolescent onset myopathy. (Available from <http://discovery.ucl.ac.uk/1528187/>).
- Porter, R.K., Scott, J.M., and Brand, M.D., 1992. Choline transport into rat liver mitochondria. *Biochemical Society Transactions*, **20**(3), p.248S.
- Price, D.C., *et al.*, 2012. *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science*, **335**(6070), pp.843-847.
- Quadros, E.V., 2010. Advances in the understanding of cobalamin assimilation and metabolism. *British Journal of Haematology*, **148**(2), pp.195-204.
- Radke, J.R., Grigera, F., Ucker, D.S., and Cook, J.L., 2014. Adenovirus E1B 19-kilodalton protein modulates innate immunity through apoptotic mimicry. *Journal of Virology*, **88**(5), pp.2658-2669.

- Ramsay, R.R., Gandour, R.D. and van der Leij, F.R., 2001. Molecular enzymology of carnitine transfer and transport. *Biochimica et Biophysica Acta*, **1546**(1), pp.21-43.
- Ramsden, D.B., Ho, P.W-L., Ho, J.W-M., Liu, H-F., So, D.H-F., Tse, H-M., Chan, K-H., and Ho, S-L., 2012. Human neuronal uncoupling proteins 4 and 5 (*UCP4* and *UCP5*): structural properties, regulation, and physiological role in protection against oxidative stress and mitochondrial dysfunction. *Brain and Behavior*, **2**(4), pp.468-478.
- Rao, A.U., Carta, L.K., Lesuisse, E., and Hamza, I., 2005. Lack of heme synthesis in a free-living eukaryote. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(12), pp.4270-4275.
- Rardin, M.J., He, W., Nishida, Y., Newman, J.C., Carrico, C., Danielson, S.R., Guo, A., Gut, P., Sahu, A.K., Li, B., Uppala, R., Fitch, M., Riff, T., Zhu, L., Zhou, J., Mulhern, D., Stevens, R.D., Ilkayeva, O.R., Newgard, C.B., Jacobson, M.P., Hellerstein, M., Goetzman, E.S., Gibson, B.W., and Verdin, E., 2013. *SIRT5* regulates the mitochondrial lysine succinylome and metabolic networks. *Cell Metabolism*, **18**(6), pp.920-933.
- Renkema, G.H., Visser, G., Baertling, F., Wintjes, L.T., Wolters, V.M., van Montfrans, J., de Kort, G.A.P., Nikkels, P.G.J., van Hasselt, P.M., van der Crabben, S.N., and Rodenburg, R.J.T., 2017. Mutated *PET117* causes complex IV deficiency and is associated with neurodevelopmental regression and medulla oblongata lesions. *Human Genetics*, **136**(6), pp.759-769.
- Rhee, S-Y., 2005. Bioinformatics. Current limitations and insights for the future. *Plant Physiology*, **138**(2), pp.569-570.
- Rhee, H-W., Zou, P., Udeshi, N.D., Martell, J.D., Mootha, V.K., Carr, S.A., and Ting, A.Y., 2013. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science*, **339**(6125), pp.1328-1331.
- Rieske, J.S., 1976. Composition, structure, and function of complex III of the respiratory chain. *Biochimica et Biophysica Acta*, **456**(2), pp.195-247.
- Rivera, M.C., Jain, R., Moore, J.E., and Lake, J.A., 1998. Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(11), pp.6239-6244.

- Rizzuto, R., Brini, M., Murgia, M., and Pozzan, T., 1993. Microdomains with high Ca^{2+} close to IP_3 -sensitive channels that are sensed by neighboring mitochondria. *Science*, **262**(5134), pp.744-747.
- Robb, S.M.C., Gotting, K., Ross, E., and Sánchez Alvarado, A., 2015. SmedGD 2.0: The *Schmidtea mediterranea* genome database. *Genesis*, **53**(8), pp.535-546.
- Roberts, D.L., Frerman, F.E., and Kim, J.J., 1996. Three-dimensional structure of human electron transfer flavoprotein to 2.1-Å resolution. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(25), pp.14355-14360.
- Robinson, A.J., Kunji, E.R.S., and Gross, A., 2012. Mitochondrial carrier homolog 2 (*MTCH2*): the recruitment and evolution of a mitochondrial carrier protein to a critical player in apoptosis. *Experimental Cell Research*, **318**(11), pp.1316-1323.
- Robinson, A.J., Overy, C., and Kunji, E.R.S., 2008. The mechanism of transport by mitochondrial carriers based on analysis of symmetry. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(46), pp.17766-17771.
- Ryan, J.F., Pang, K., Schnitzler, C.E., Nguyen, A-D., Moreland, R.T., Simmons, D.K., Koch, B.J., Francis, W.R., Havlak, P., NISC Comparative Sequencing Program, Smith, S.A., Putnam, N.H., Haddock, S.H.D., Dunn, C.W., Wolfsberg, T.G., Mullikin, J.C., Martindale, M.Q., and Baxevanis, A.D., 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science*, **342**(6164), p.1242592.
- Ryter, S.W. and Tyrrell, R.M., 2000. The heme synthesis and degradation pathways: role in oxidant sensitivity. Heme oxygenase has both pro- and anti-oxidant properties. *Free Radical Biology & Medicine*, **28**(2), pp.289-309.
- Sacksteder, K.A., Morrell, J.C., Wanders, R.J.A., Matalon, R., and Gould, S.J., 1999. MCD encodes peroxisomal and cytoplasmic forms of malonyl-CoA decarboxylase and is mutated in malonyl-CoA decarboxylase deficiency. *The Journal of Biological Chemistry*, **274**(35), pp.24461-24468.
- Sagan, L., 1967. On the origin of mitosing cells. *Journal of Theoretical Biology*, **14**(3), pp.255-274.

- Saisawat, P. *et al.*, 2014. Whole-exome resequencing reveals recessive mutations in *TRAP1* in individuals with CAKUT and VACTERL association. *Kidney International*, **85**(6), pp.1310-1317.
- Saleheen, D. *et al.*, 2017. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature*, **544**(7649), pp.235-239.
- Santana, M., Ionescu, M.S., Vertes, A., Longin, R., Kunst, F., Danchin, A., and Glaser, P., 1994. *Bacillus subtilis* F₀F₁ ATPase: DNA sequence of the *atp* operon and characterization of *atp* mutants. *Journal of Bacteriology*, **176**(22), pp.6802-6811.
- Sass, J.O., Fischer, K., Wang, R., Christensen, E., Scholl-Bürgi, S., Chang, R., Kapelari, K., and Walter, M., 2010. *D*-glyceric aciduria is caused by genetic deficiency of *D*-glycerate kinase (*GLYCK*). *Human Mutation*, **31**(12), pp.1280-1285.
- Satoh, A., Brace, C.S., Rensing, N., Cliften, P., Wozniak, D.F., Herzog, E.D., Tamada, K.A., and Imai, S-I., 2013. *Sirt1* extends life span and delays aging in mice through the regulation of Nk2 homeobox 1 in the DMH and LH. *Cell Metabolism*, **18**(3), pp.416-430.
- Sánchez-Blanco, A., Fridell, Y-W.C., and Helfand, S.L., 2006. Involvement of *Drosophila* uncoupling protein 5 in metabolism and aging. *Genetics*, **172**(3), pp.1699-1710.
- Schaecher, S.R., Touchette, E., Schriewer, J., Buller, R.M., and Pekosz, A., 2007. Severe acute respiratory syndrome coronavirus gene 7 products contribute to virus-induced apoptosis. *Journal of Virology*, **81**(20), pp.11054-11068.
- Schägger, H. and Pfeiffer, K., 2000. Supercomplexes in the respiratory chains of yeast and mammalian mitochondria. *The EMBO Journal*, **19**(8), pp.1777-1783.
- Schwartz, M.S., Benci, J.L., Selote, D.S., Sharma, A.K., Chen, A.G.Y., Dang, H., Fares, H., and Vatamaniuk, O.K., 2010. Detoxification of multiple heavy metals by a half-molecule ABC transporter, *HMT-1*, and coelomocytes of *Caenorhabditis elegans*. *PloS One*, **5**(3), e9564.
- Schwarz, G., Mendel, R.R., and Ribbe, M.W., 2009. Molybdenum cofactors, enzymes and pathways. *Nature*, **460**(7257), pp.839-847.

- Schwer, B., Eckersdorff, M., Li, Y., Silva, J.C., Fermin, D., Kurtev, M.V., Giallourakis, C., Comb, M.J., Alt, F.W., and Lombard, D.B., 2009. Calorie restriction alters mitochondrial protein acetylation. *Aging Cell*, **8**(5), pp.604-606.
- Serre, V., Rozanska, A., Beinat, M., Chretien, D., Boddaert, N., Munnich, A., Rötig, A., and Chrzanowska-Lightowlers, Z.M., 2013. Mutations in mitochondrial ribosomal protein *MRPL12* leads to growth retardation, neurological deterioration and mitochondrial translation deficiency. *Biochimica et Biophysica Acta*, **1832**(8), pp.1304-1312.
- Shamseldin, H.E., Smith, L.L., Alkhalidi, H., Summers, B., Alsedairy, H., Xiong, Y., Gupta, V.A., and Alkuraya, F.S., 2016. Mutation of the mitochondrial carrier *SLC25A42* causes a novel form of mitochondrial myopathy in humans. *Human Genetics*, **135**(1), pp.21-30.
- Sheftel, A.D., Stehling, O., Pierik, A.J., Netz, D.J.A., Kerscher, S., Elsässer, H-P., Wittig, I., Balk, J., Brandt, U., and Lill, R., 2009. Human *ind1*, an iron-sulfur cluster assembly factor for respiratory complex I. *Molecular and Cellular Biology*, **29**(22), pp.6059-6073.
- Sherman, P.W. and Jarvis, J.U.M., 2002. Extraordinary life spans of naked mole-rats (*Heterocephalus glaber*). *Journal of Zoology*, **258**(3), pp.307-311.
- Shoguchi, E., *et al.*, 2013. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Current Biology*, **23**(15), pp.1399-1408.
- Shumar, S.A., Kerr, E.W., Geldenhuys, W.J., Montgomery, G.E., Fagone, P., Thirawatananond, P., Saavedra, H., Gabelli, S.B., and Leonardi, R., 2018. *Nudt19* is a renal CoA diphosphohydrolase with biochemical and regulatory properties that are distinct from the hepatic *Nudt7* isoform. *The Journal of Biological Chemistry*, **293**(11), pp.4134-4148.
- Sibbing, D., *et al.*, 2011. Mutations in the mitochondrial thioredoxin reductase gene *TXNRD2* cause dilated cardiomyopathy. *European Heart Journal*, **32**(9), pp.1121-1133.
- Skladal, D., Halliday, J., and Thorburn, D.R., 2003. Minimum birth prevalence of mitochondrial respiratory chain disorders in children. *Brain*, **126**(8), pp.1905-1912.
- Smith, A.C. and Robinson, A.J., 2016. MitoMiner v3.1, an update on the mitochondrial proteomics database. *Nucleic Acids Research*, **44**(D1), pp.D1258-D1261.

- Smith, A.C., Eyassu, F., Mazat, J-P., and Robinson, A.J., 2017. MitoCore: a curated constraint-based model for simulating human central metabolism. *BMC Systems Biology*, **11**, p.114.
- Smith, D.G.S., Gawryluk, R.M.R, Spencer, D.F., Pearlman, R.E., Siu, K.W.M., and Gray, M.W., 2007. Exploring the mitochondrial proteome of the ciliate protozoon *Tetrahymena thermophila*: direct analysis by tandem mass spectrometry. *Journal of Molecular Biology*, **374**(3), pp.837-863.
- Smith, P.M., Fox, J.L., and Winge, D.R., 2012. Biogenesis of the cytochrome *bc₁* complex and role of assembly factors. *Biochimica et Biophysica Acta*, **1817**(2), pp.276-286.
- Snitkin, E.S., Gustafson, A.M., Mellor, J., Wu, J., and DeLisi, C., 2006. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics*, **7**, p.420.
- Sofou, K., Kollberg, G., Holmström, M., Dávila, M., Darin, N., Gustafsson, C.M., Holme, E., Oldfors, A., Tulinius, M., and Asin-Cayuela, J., 2015. Whole exome sequencing reveals mutations in *NARS2* and *PARS2*, encoding the mitochondrial asparaginyl-tRNA synthetase and prolyl-tRNA synthetase, in patients with Alpers syndrome. *Molecular Genetics & Genomic Medicine*, **3**(1), pp.59-68.
- Sorgato, M.C., Moran, O., and Pedersen, P.L., 1993. Channels in mitochondrial membranes: knowns, unknowns, and prospects for the future. *Critical Reviews in Biochemistry and Molecular Biology*, **28**(2), pp.127-171.
- Spelbrink, J.N. *et al.*, 2001. Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. *Nature Genetics*, **28**(3), pp.223-231.
- Spiegel, R., Saada, A., Halvardson, J., Soifermann, D., Shaag, A., Edvardson, S., Horovitz, Y., Khayat, M., Shalev, S.A., Feuk, L., and Elpeleg, O., 2013. Deleterious mutation in *FDX1L* gene is associated with a novel mitochondrial muscle myopathy. *European Journal of Human Genetics*, **22**(7), pp.902-906.

- Stehling, O. and Lill, R., 2013. The role of mitochondria in cellular iron-sulfur protein biogenesis: mechanisms, connected processes, and diseases. *Cold Spring Harbor Perspectives in Biology*, **5**(8), a011312.
- Strecker, H.J., 1965. Purification and properties of rat liver ornithine δ -transaminase. *The Journal of Biological Chemistry*, **240**, pp.1225-1230.
- Strijbis, K., Vaz, F.M., and Distel, B., 2010. Enzymology of the carnitine biosynthesis pathway. *IUBMB Life*, **62**(5), pp.357-362.
- Strittmatter, L., Li, Y., Nakatsuka, N.J., Calvo, S.E., Grabarek, Z., and Mootha, V.K., 2014. *CLYBL* is a polymorphic human enzyme with malate synthase and β -methylmalate synthase activity. *Human Molecular Genetics*, **23**(9), pp.2313-2323.
- Suga, H., Chen, Z., de Mendoza, A., Seb -Pedr s, A., Brown, M.W., Kramer, E., Carr, M., Kerner, P., Vervoort, M., S nchez-Pons, N., Torruella, G., Derelle, R., Manning, G., Lang, B.F., Russ, C., Haas, B.J., Roger, A.J., Nusbaum, C., and Ruiz-Trillo, I., 2013. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. *Nature Communications*, **4**, p.2325.
- Sulem, P., Helgason, H., Oddson, A., Stefansson, H., Gudjonsson, S.A., Zink, F., Hjartarson, E., Sigurdsson, G.T., Jonasdottir, A., Jonasdottir, A., Sigurdsson, A., Magnusson, O.T., Kong, A., Helgason, A., Holm, H., Thorsteinsdottir, T., Masson, G., Gudbjartsson, D.F., and Stefansson, K., 2015. Identification of a large set of rare complete human knockouts. *Nature Genetics*, **47**(5), pp.448-452.
- Sun, F., Huo, X., Zhai, Y., Wang, A., Xu, J., Su, D., Bartlam, M., and Rao, Z., 2005. Crystal structure of mitochondrial respiratory membrane protein complex II. *Cell*, **121**(7), pp.1043-1057.
- Sun, S. and Habermann, B.H., 2017. A guide to computational methods for predicting mitochondrial localization. *Methods in Molecular Biology*, **1567**(1), pp.1–14.
- Sutak, R., Dolezal, P., Fiumera, H.L., Hrdy, I., Dancis, A., Delgadillo-Correa, M., Johnson, P.J., M ller, M., and Tachezy, J., 2004. Mitochondrial-type assembly of FeS centers in the hydrogenosomes of the amitochondriate eukaryote *Trichomonas vaginalis*. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(28), pp.10368-10373.

- Suzan-Monti, M., La Scola, B., Barrassi, L., Espinosa, L., and Raoult, D., 2007. Ultrastructural characterization of the giant volcano-like virus factory of *Acanthamoeba polyphaga* Mimivirus. *PloS One*, **2**(3), e328.
- Svinkina, T., Gu, H., Silva, J.C., Mertins, P., Qiao, J., Fereshetian, S., Jaffe, J.D., Kuhn, E., Udeshi, N.D., and Carr, S.A., 2015. Deep, quantitative coverage of the lysine acetylome using novel anti-acetyl-lysine antibodies and an optimized proteomic workflow. *Molecular & Cellular Proteomics*, **14**(9), pp.2429-2440.
- Swanson, M.A., Garcia, S.M., Spector, E., Kronquist, K., Creadon-Swindell, G., Walter, M., Christensen, E., van Hove, J.L.K., and Sass, J.O., 2017. D-glyceric aciduria does not cause non-ketotic hyperglycinemia: a historic co-occurrence. *Molecular Genetics and Metabolism*, **121**(2), pp.80-82.
- Szklarczyk, R. and Huynen, M.A., 2010. Mosaic origin of the mitochondrial proteome. *Proteomics*, **10**(22), pp.4012-4024.
- Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., Costa, J., Fraifeld, V.E., and de Magalhães, J.P., 2013. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Research*, **41**(D1), pp.D1027-D1033.
- Tatusov, R.L., Koonin, E.V., and Lipman, D.J., 1997. A genomic perspective on protein families. *Science*, **278**(5338), pp.631-637.
- Taylor, R.W. *et al.*, 2014. Use of whole-exome sequencing to determine the genetic basis of multiple mitochondrial respiratory chain complex deficiencies. *JAMA*, **312**(1), pp.68-77.
- Teschner, J., Lachmann, N., Schulze, J., Geisler, M., Selbach, K., Santamaria-Araujo, J., Balk, J., Mendel, R.R., and Bittner, F., 2010. A novel role for *Arabidopsis* mitochondrial ABC transporter *ATM3* in molybdenum cofactor biosynthesis. *The Plant Cell*, **22**(2), pp.468-480.
- The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **45**(D1), pp.D158-D169.

- Thomas, P.D., Wood, V., Mungall, C.J., Lewis, S.E., Blake, J.A., and Gene Ontology Consortium, 2012. On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: A short report. *PLoS Computational Biology*, **8**(2), e1002386.
- Thompson, J.D., Gibson, T.J., and Higgins, D.G., 2002. Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, Chapter 2, Unit 2.3.
- Thul, P.J. *et al.*, 2017. A subcellular map of the human proteome. *Science*, **356**(6340), eaal3321.
- Titus, S.A. and Moran, R.G., 2000. Retrovirally mediated complementation of the glyB phenotype. Cloning of a human gene encoding the carrier for entry of folates into mitochondria. *The Journal of Biological Chemistry*, **275**(47), pp.36811-36817.
- Tondera, D., Santel, A., Schwarzer, R., Dames, S., Giese, K., Klippel, A., and Kaufmann, J., 2004. Knockdown of *MTP18*, a novel phosphatidylinositol 3-kinase-dependent protein, affects mitochondrial morphology and induces apoptosis. *The Journal of Biological Chemistry*, **279**(30), pp.31544-31555.
- Tondera, D., Czauderna, F., Paulick, K., Schwarzer, R., Kaufmann, J., and Santel, A., 2005. The mitochondrial protein *MTP18* contributes to mitochondrial fission in mammalian cells. *Journal of Cell Science*, **118**(14), pp.3049-3059.
- Tuzovic, L., Yu, L., Zeng, W., Li, X., Lu, H., Lu, H-M., Gonzalez, K.D.F., and Chung, W.K., 2013. A human *de novo* mutation in *MYH10* phenocopies the loss of function mutation in mice. *Rare Diseases*, **1**, e26144.
- Tye, B.K., Chien, J., Lehman, I.R., Duncan, B.K., and Warner, H.R., 1978. Uracil incorporation: a source of pulse-labeled DNA fragments in the replication of the *Escherichia coli* chromosome. *Proceedings of the National Academy of Sciences of the United States of America*, **75**(1), pp.233-237.
- Uhlén, M., *et al.*, 2015. Tissue-based map of the human proteome. *Science*, **347**(6220), p.1260419.

- Van Goethem, G., Löfgren, A., Dermaut, B., Ceuterick, C., Marin, J-J., and van Broeckhoven, C., 2003. Digenic progressive external ophthalmoplegia in a sporadic patient: recessive mutations in *POLG* and *C10orf2*/Twinkle. *Human Mutation*, **22**(2), pp.175-176.
- Van Haute, L., Dietmann, S., Kremer, L., Hussain, S., Pearce, S.F., Powell, C.A., Rorbach, J., Lantaff, R., Blanco, S., Sauer, S., Kotzaeridou, U., Hoffmann, G.F., Memari, Y., Kolb-Kokocinski, A., Durbin, R., Mayr, J.A., Frye, M., Prokisch, H., and Minczuk, M., 2016. Deficient methylation and formylation of mt-tRNA(Met) wobble cytosine in a patient carrying mutations in *NSUN3*. *Nature Communications*, **7**, p.12039.
- Van Schaftingen, E., Rzem, R., and Veiga-da-Cunha, M., 2009. L-2-Hydroxyglutaric aciduria, a disorder of metabolite repair. *Journal of Inherited Metabolic Disease*, **32**(2), pp.135-142.
- Vávra, J., 2005. "Polar vesicles" of microsporidia are mitochondrial remnants ("mitosomes")? *Folia Parasitologica*, **52**(1-2), pp.193-195.
- Vinothkumar, K.R., Zhu, J., and Hirst, J., 2014. Architecture of mammalian respiratory complex I. *Nature*, **515**(7525), pp.80-84.
- Wagner, G.R. and Hirschey, M.D., 2014. Nonenzymatic protein acylation as a carbon stress regulated by sirtuin deacylases. *Molecular Cell*, **54**(1), pp.5-16.
- Wagner, G.R. and Payne, R.M., 2013. Widespread and enzyme-independent *N*^ε-acetylation and *N*^ε-succinylation of proteins in the chemical conditions of the mitochondrial matrix. *The Journal of Biological Chemistry*, **288**(40), pp.29036-29045.
- Walker, J.E. and Runswick, M.J., 1993. The mitochondrial transport protein superfamily. *Journal of Bioenergetics and Biomembranes*, **25**(5), pp.435-446.
- Wallin, I.E., 1927. *Symbiontism and the Origin of Species*, Baltimore: Williams & Wilkins Company.
- Wanders, R.J., Romeyn, G.J., van Roermund, C.W.T., Schutgens, R.B.H., van den Bosch, H., and Tager, J.M., 1988. Identification of L-pipecolate oxidase in human liver and its deficiency in the Zellweger syndrome. *Biochemical and Biophysical Research Communications*, **154**(1), pp.33-38.

- Wanders, R.J., Romeyn, G.J., Schutgens, R.B.H., and Tager, J.M., 1989. *L*-pipecolate oxidase: a distinct peroxisomal enzyme in man. *Biochemical and Biophysical Research Communications*, **164**(1), pp.550-555.
- Wang, C. and Youle, R.J., 2009. The role of mitochondria in apoptosis. *Annual Review of Genetics*, **43**(1), pp.95-118.
- Wang, Z. and Wu, M., 2014. Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite. *PloS One*, **9**(10), e110685.
- Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J., 2009. Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9), pp.1189-1191.
- Watt, I.N., Montgomery, M.G., Runswick, M.J., Leslie, A.G.W., and Walker, J.E., 2010. Bioenergetic cost of making an adenosine triphosphate molecule in animal mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(39), pp.16823-16827.
- Webb, B. and Sali, A., 2017. Comparative Protein Structure Modeling Using MODELLER. *Methods in Molecular Biology*, **1654**, pp.39-54.
- Weinert, B.T. *et al.*, 2014. Acetylation dynamics and stoichiometry in *Saccharomyces cerevisiae*. *Molecular Systems Biology*, **10**, p.716.
- Weinert, B.T., Iesmantavicius, V., Moustafa, T., Schölz, C., Wagner, S.A., Magnes, C., Zechner, R., and Choudhary, C., 2015. Analysis of acetylation stoichiometry suggests that *SIRT3* repairs non-enzymatic acetylation lesions. *The EMBO Journal*, **34**(21), pp.2620-2632.
- Weinert, B.T., Wagner, S.A., Horn, H., Henriksen, P., Liu, W.R., Olsen, J.V., Jensen, L.J., and Choudhary, C., 2011. Proteome-wide mapping of the *Drosophila* acetylome demonstrates a high degree of conservation of lysine acetylation. *Science Signaling*, **4**(183), ra48.
- Westermann, B., 2010. Mitochondrial fusion and fission in cell life and death. *Nature Reviews: Molecular Cell Biology*, **11**(12), pp.872-884.
- Weynberg, K.D., Allen, M.J., and Wilson, W.H., 2017. Marine prasinoviruses and their tiny plankton hosts: a review. *Viruses*, **9**(3), p.43.

- White, E., Cipriani, R., Sabbatini, P., and Denton, A., 1991. Adenovirus E1B 19-kilodalton protein overcomes the cytotoxicity of E1A proteins. *Journal of Virology*, **65**(6), pp.2968-2978.
- Wilkinson, G.S. and South, J.M., 2002. Life history, ecology and longevity in bats. *Aging Cell*, **1**(2), pp.124-131.
- Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., McNicholas, S.J., Murshudoc, G.N., Pannu, N.S., Potterton, E.A., Powell, H.R., Read, R.J., Vagin, A., and Wilson, K.S., 2011. Overview of the CCP4 suite and current developments. *Acta Crystallographica. Section D, Biological Crystallography*, **67**(4), pp.235-242.
- Wnęk, M., Ressel, L., Ricci, E., Rodriguez-Martinez, C., Guerrero, J.C.V., Ismail, Z., Smith, C., Kipar, A., Sodeik, B., Chinnery, P.F., Solomon, T., and Griffiths, M.J., 2016. Herpes simplex encephalitis is linked with selective mitochondrial damage; a post-mortem and in vitro study. *Acta Neuropathologica*, **132**(3), pp.433-451.
- Yamauchi, A., Uchida, S., Kwon, H.M., Preton, A.S., Robey, R.B., Garcia-Perez, A., Burg, M.B., and Handler, J.S., 1992. Cloning of a Na⁺- and Cl⁻-dependent betaine transporter that is regulated by hypertonicity. *The Journal of Biological Chemistry*, **267**(1), pp.649–652.
- Yang, Y., Xiong, J., Zhou, Z., Huo, F., Miao, W., Ran, C., Liu, Y., Zhang, J., Feng, J., Wang, M.W.M., Wang, L., and Yao, B., 2014. The genome of the myxosporean *Thelohanellus kitauei* shows adaptations to nutrient acquisition within its fish host. *Genome Biology and Evolution*, **6**(12), pp.3182-3198.
- Yates, A. *et al.*, 2016. Ensembl 2016. *Nucleic Acids Research*, **44**(D1), pp.D710-D716.
- Yoshida, M., Muneyuki, E., and Hisabori, T., 2001. ATP synthase – a marvellous rotary engine of the cell. *Nature Reviews: Molecular Cell Biology*, **2**(9), pp.669-677.
- Yu, X.X., Mao, W., Zhong, A., Schow, P., Brush, J., Sherwood, S.W., Adams, S.H., and Pan, G., 2000. Characterization of novel *UCP5/BMCP1* isoforms and differential regulation of *UCP4* and *UCP5* expression through dietary or temperature manipulation. *FASEB Journal*, **14**(11), pp.1611-1618.

- Zhang, G. *et al.*, 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346**(6215), pp.1311-1320.
- Zhang, Y., Rodionov, D.A., Gelfand, M.S., and Gladyshev, V.N., 2009. Comparative genomic analyses of nickel, cobalt and vitamin B12 utilization. *BMC Genomics*, **10**, p.78.
- Zhao, F., Wu, J., Xue, A., Su, Y., Wang, X., Lu, X., Zhou, Z., Qu, J., and Zhou, X., 2013. Exome sequencing reveals *CCDC111* mutation associated with high myopia. *Human Genetics*, **132**(8), pp.913-921.
- Zhou, J., Richardson, A.J., and Rudd, K.E., 2013. EcoGene-RefSeq: EcoGene tools applied to the RefSeq prokaryotic genomes. *Bioinformatics*, **29**(15), pp.1917-1918.
- Zhu, J., Vinothkumar, K.R., and Hirst, J., 2016. Structure of mammalian respiratory complex I. *Nature*, **536**(7616), pp.354-358.
- Zick, M., Rabl, R., and Reichert, A.S., 2009. Cristae formation-linking ultrastructure and function of mitochondria. *Biochimica et Biophysica Acta*, **1793**(1), pp.5-19.
- Zou, H., Li, Y., Liu, X., and Wang, X., 1999. An *APAF-1*.cytochrome *c* multimeric complex is a functional apoptosome that activates procaspase-9. *The Journal of Biological Chemistry*, **274**(17), pp.11549-11556.

Appendices

Appendix I: Orthology dataset summary

Table 1. Species included in the orthology dataset with number of predicted IMPI 2017 orthologues.

Species	Group	No. orthologues
<i>Homo sapiens</i>	Holozoa	1,550
<i>Pan troglodytes</i>	Holozoa	1,531
<i>Gorilla gorilla</i>	Holozoa	1,486
<i>Pongo abelii</i>	Holozoa	1,488
<i>Nomascus leucogenys</i>	Holozoa	1,489
<i>Macaca fascicularis</i>	Holozoa	1,534
<i>Callithrix jacchus</i>	Holozoa	1,514
<i>Saimiri boliviensis</i>	Holozoa	1,501
<i>Carlito syrichta</i>	Holozoa	1,459
<i>Otolemur garnettii</i>	Holozoa	1,461
<i>Mus musculus</i>	Holozoa	1,504
<i>Rattus norvegicus</i>	Holozoa	1,504
<i>Cricetulus griseus</i>	Holozoa	1,498
<i>Cavia porcellus</i>	Holozoa	1,485
<i>Heterocephalus glaber</i>	Holozoa	1,508
<i>Oryctolagus cuniculus</i>	Holozoa	1,429
<i>Bos taurus</i>	Holozoa	1,510
<i>Equus caballus</i>	Holozoa	1,472
<i>Canis familiaris</i>	Holozoa	1,444
<i>Odobenus rosmarus</i>	Holozoa	1,497
<i>Orcinus orca</i>	Holozoa	1,489
<i>Myotis lucifugus</i>	Holozoa	1,402
<i>Loxodonta africana</i>	Holozoa	1,400
<i>Dasypus novemcinctus</i>	Holozoa	1,457
<i>Sarcophilus harrisii</i>	Holozoa	1,347
<i>Monodelphis domestica</i>	Holozoa	1,429
<i>Ornithorhynchus anatinus</i>	Holozoa	1,233
<i>Melopsittacus undulatus</i>	Holozoa	1,190
<i>Gallus gallus</i>	Holozoa	1,286
<i>Alligator mississippiensis</i>	Holozoa	1,371
<i>Python bivittatus</i>	Holozoa	1,361
<i>Anolis carolinensis</i>	Holozoa	1,355
<i>Chrysemys picta</i>	Holozoa	1,439
<i>Chelonia mydas</i>	Holozoa	1,377
<i>Nanorana parkeri</i>	Holozoa	1,330
<i>Xenopus tropicalis</i>	Holozoa	1,365
<i>Latimeria chalumnae</i>	Holozoa	1,352

Species	Group	No. orthologues
<i>Oryzias latipes</i>	Holozoa	1,317
<i>Oreochromis niloticus</i>	Holozoa	1,350
<i>Esox lucius</i>	Holozoa	1,366
<i>Danio rerio</i>	Holozoa	1,357
<i>Scleropages formosus</i>	Holozoa	1,362
<i>Lepisosteus oculatus</i>	Holozoa	1,287
<i>Callorhinchus milii</i>	Holozoa	1,343
<i>Oikopleura dioica</i>	Holozoa	718
<i>Ciona intestinalis</i>	Holozoa	909
<i>Branchiostoma floridae</i>	Holozoa	1,024
<i>Strongylocentrotus purpuratus</i>	Holozoa	1,088
<i>Saccoglossus kowalevskii</i>	Holozoa	1,056
<i>Acyrtosiphon pisum</i>	Holozoa	916
<i>Pediculus humanus corporis</i>	Holozoa	922
<i>Danaus plexippus</i>	Holozoa	903
<i>Bombyx mori</i>	Holozoa	927
<i>Aedes aegypti</i>	Holozoa	946
<i>Anopheles gambiae</i>	Holozoa	945
<i>Drosophila melanogaster</i>	Holozoa	946
<i>Camponotus floridanus</i>	Holozoa	906
<i>Apis mellifera</i>	Holozoa	930
<i>Tribolium castaneum</i>	Holozoa	983
<i>Orchesella cincta</i>	Holozoa	864
<i>Daphnia pulex</i>	Holozoa	964
<i>Hyalella azteca</i>	Holozoa	953
<i>Limulus polyphemus</i>	Holozoa	1,014
<i>Ixodes scapularis</i>	Holozoa	895
<i>Priapulus caudatus</i>	Holozoa	990
<i>Caenorhabditis elegans</i>	Holozoa	846
<i>Caenorhabditis briggsae</i>	Holozoa	830
<i>Strongyloides ratti</i>	Holozoa	725
<i>Ascaris suum</i>	Holozoa	675
<i>Brugia malayi</i>	Holozoa	749
<i>Loa loa</i>	Holozoa	732
<i>Trichinella spiralis</i>	Holozoa	614
<i>Trichuris suis</i>	Holozoa	640
<i>Schmidtea mediterranea</i>	Holozoa	733
<i>Clonorchis sinensis</i>	Holozoa	724
<i>Opisthorchis viverrini</i>	Holozoa	720
<i>Schistosoma mansoni</i>	Holozoa	709
<i>Schistosoma haematobium</i>	Holozoa	663
<i>Hymenolepis microstoma</i>	Holozoa	659

Species	Group	No. orthologues
<i>Echinococcus granulosus</i>	Holozoa	715
<i>Echinococcus multilocularis</i>	Holozoa	704
<i>Aplysia californica</i>	Holozoa	1,042
<i>Lottia gigantea</i>	Holozoa	1,010
<i>Crassostrea gigas</i>	Holozoa	858
<i>Helobdella robusta</i>	Holozoa	918
<i>Capitella teleta</i>	Holozoa	1,046
<i>Mnemiopsis leidyi</i>	Holozoa	675
<i>Nematostella vectensis</i>	Holozoa	961
<i>Acropora digitifera</i>	Holozoa	908
<i>Hydra vulgaris</i>	Holozoa	825
<i>Thelohanellus kitauei</i>	Holozoa	228
<i>Amphimedon queenslandica</i>	Holozoa	855
<i>Trichoplax adhaerens</i>	Holozoa	885
<i>Monosiga brevicollis MX1</i>	Holozoa	604
<i>Salpingoeca rosetta</i>	Holozoa	713
<i>Sphaeroforma arctica</i>	Holozoa	639
<i>Capsaspora owczarzaki ATCC 30864</i>	Holozoa	760
<i>Leptosphaeria maculans JN3</i>	Holomycota	574
<i>Aspergillus nidulans FGSC A4</i>	Holomycota	584
<i>Neurospora crassa OR74A</i>	Holomycota	596
<i>Botrytis cinerea T4</i>	Holomycota	572
<i>Xylona heveae TC161</i>	Holomycota	584
<i>Tuber melanosporum Mel28</i>	Holomycota	555
<i>Dactylellina haptotyla CBS 200.50</i>	Holomycota	566
<i>Debaryomyces hansenii CBS767</i>	Holomycota	535
<i>Scheffersomyces stipitis CBS 6054</i>	Holomycota	509
<i>Saccharomyces cerevisiae S288c</i>	Holomycota	478
<i>Candida glabrata CBS 138</i>	Holomycota	479
<i>Candida albicans SC5314</i>	Holomycota	493
<i>Yarrowia lipolytica CLIB122</i>	Holomycota	559
<i>Schizosaccharomyces pombe 972h-</i>	Holomycota	500
<i>Pneumocystis jirovecii RU7</i>	Holomycota	432
<i>Coprinopsis cinerea</i>	Holomycota	589
<i>Serpula lacrymans var. lacrymans S7.9</i>	Holomycota	548
<i>Calocera cornea HHB12733</i>	Holomycota	568
<i>Trichosporon asahii var. asahii CBS 2479</i>	Holomycota	502
<i>Cryptococcus gattii WM276</i>	Holomycota	542
<i>Wallemia mellicola CBS 633.66</i>	Holomycota	529
<i>Tilletiaria anomala UBC 951</i>	Holomycota	536
<i>Ustilago maydis 521</i>	Holomycota	550
<i>Melampsora larici-populina 98AG31</i>	Holomycota	541

Species	Group	No. orthologues
<i>Puccinia graminis</i> f. sp. <i>tritici</i> CRL 75-36-700-3	Holomycota	519
<i>Rhodotorula graminis</i> WP1	Holomycota	560
<i>Mixia osmundae</i> IAM 14324	Holomycota	554
<i>Rhizophagus irregularis</i> DAOM 181602	Holomycota	647
<i>Mucor circinelloides</i> f. <i>circinelloides</i> 1006PhL	Holomycota	675
<i>Rhizopus delemar</i> RA 99-880	Holomycota	612
<i>Mortierella verticillata</i> NRRL 6337	Holomycota	677
<i>Allomyces macrogynus</i> ATCC 38327	Holomycota	621
<i>Spizellomyces punctatus</i> DAOM BR117	Holomycota	645
<i>Batrachochytrium dendrobatidis</i> JAM81	Holomycota	589
<i>Gonapodya prolifera</i> JEL478	Holomycota	612
<i>Nematocida parisii</i> ERTm3	Holomycota	104
<i>Edhazardia aedis</i> USNM 41457	Holomycota	114
<i>Vavraia culicis</i>	Holomycota	118
<i>Trachipleistophora hominis</i>	Holomycota	119
<i>Spraguea lophii</i> 42_110	Holomycota	104
<i>Encephalitozoon intestinalis</i> ATCC 50506	Holomycota	111
<i>Encephalitozoon cuniculi</i> GB-M1	Holomycota	111
<i>Enterocytozoon bieneusi</i> H348	Holomycota	73
<i>Rozella allomycis</i> CSF55	Holomycota	424
<i>Fonticula alba</i>	Holomycota	511
<i>Polysphondylium pallidum</i> PN500	Amoebozoa	556
<i>Dictyostelium purpureum</i>	Amoebozoa	553
<i>Dictyostelium discoideum</i> AX4	Amoebozoa	567
<i>Dictyostelium fasciculatum</i>	Amoebozoa	569
<i>Acytostelium subglobosum</i> LB1	Amoebozoa	577
<i>Acanthamoeba castellanii</i> str. Neff	Amoebozoa	560
<i>Entamoeba dispar</i> SAW760	Amoebozoa	126
<i>Entamoeba histolytica</i> HM-1:IMSS	Amoebozoa	137
<i>Nicotiana tabacum</i>	Archaeplastida	600
<i>Solanum lycopersicum</i>	Archaeplastida	589
<i>Beta vulgaris</i> subsp. <i>vulgaris</i>	Archaeplastida	589
<i>Cucumis sativus</i>	Archaeplastida	591
<i>Fragaria vesca</i> subsp. <i>vesca</i>	Archaeplastida	595
<i>Arabidopsis lyrata</i> subsp. <i>lyrata</i>	Archaeplastida	571
<i>Arabidopsis thaliana</i>	Archaeplastida	592
<i>Vitis vinifera</i>	Archaeplastida	599
<i>Nelumbo nucifera</i>	Archaeplastida	596
<i>Musa acuminata</i> subsp. <i>malaccensis</i>	Archaeplastida	592
<i>Sorghum bicolor</i>	Archaeplastida	561
<i>Brachypodium distachyon</i>	Archaeplastida	582
<i>Oryza sativa</i> Japonica Group	Archaeplastida	596

Species	Group	No. orthologues
<i>Zea mays</i>	Archaeplastida	596
<i>Elaeis guineensis</i>	Archaeplastida	582
<i>Zostera marina</i>	Archaeplastida	571
<i>Amborella trichopoda</i>	Archaeplastida	594
<i>Selaginella moellendorffii</i>	Archaeplastida	573
<i>Physcomitrella patens</i>	Archaeplastida	598
<i>Marchantia polymorpha</i> subsp. <i>polymorpha</i>	Archaeplastida	545
<i>Chlorella variabilis</i>	Archaeplastida	542
<i>Coccomyxa subellipsoidea</i> C-169	Archaeplastida	552
<i>Chlamydomonas reinhardtii</i>	Archaeplastida	526
<i>Volvox carteri</i> f. <i>nagariensis</i>	Archaeplastida	513
<i>Gonium pectorale</i>	Archaeplastida	505
<i>Ostreococcus tauri</i>	Archaeplastida	501
<i>Ostreococcus lucimarinus</i> CCE9901	Archaeplastida	481
<i>Micromonas pusilla</i> CCMP1545	Archaeplastida	508
<i>Klebsormidium nitens</i>	Archaeplastida	607
<i>Galdieria sulphuraria</i>	Archaeplastida	518
<i>Chondrus crispus</i>	Archaeplastida	435
<i>Cyanidioschyzon merolae</i> strain 10D	Archaeplastida	490
<i>Cyanophora paradoxa</i>	Archaeplastida	440
<i>Vitrella brassicaformis</i> CCMP3155	SAR Group	521
<i>Oxytricha trifallax</i>	SAR Group	455
<i>Stylonychia lemnae</i>	SAR Group	428
<i>Paramecium tetraurelia</i> strain d4-2	SAR Group	422
<i>Tetrahymena thermophila</i> SB210	SAR Group	454
<i>Ichthyophthirius multifiliis</i>	SAR Group	363
<i>Symbiodinium minutum</i> Clade B1	SAR Group	413
<i>Perkinsus marinus</i> ATCC 50983	SAR Group	410
<i>Reticulomyxa filosa</i>	SAR Group	397
<i>Ectocarpus siliculosus</i>	SAR Group	586
<i>Aureococcus anophagefferens</i>	SAR Group	461
<i>Nannochloropsis gaditana</i>	SAR Group	566
<i>Blastocystis hominis</i>	SAR Group	339
<i>Thalassiosira oceanica</i>	SAR Group	465
<i>Thalassiosira pseudonana</i> CCMP1335	SAR Group	497
<i>Phaeodactylum tricornutum</i>	SAR Group	534
<i>Aphanomyces invadans</i>	SAR Group	635
<i>Saprolegnia diclina</i> VS20	SAR Group	621
<i>Albugo laibachii</i> Nc14	SAR Group	513
<i>Albugo candida</i>	SAR Group	511
<i>Phytophthora infestans</i> T30-4	SAR Group	593
<i>Phytophthora sojae</i>	SAR Group	609

Species	Group	No. orthologues
<i>Babesia bovis</i> T2Bo	SAR Group	280
<i>Eimeria tenella</i>	SAR Group	309
<i>Eimeria maxima</i>	SAR Group	206
<i>Eimeria acervulina</i>	SAR Group	246
<i>Cryptosporidium hominis</i> TU502	SAR Group	164
<i>Cryptosporidium muris</i> RN66	SAR Group	205
<i>Cryptosporidium parvum</i> Iowa II	SAR Group	179
<i>Theileria annulata</i> strain Ankara	SAR Group	273
<i>Theileria parva</i>	SAR Group	270
<i>Theileria orientalis</i> strain Shintoku	SAR Group	275
<i>Toxoplasma gondii</i> GT1	SAR Group	400
<i>Neospora caninum</i> Liverpool	SAR Group	383
<i>Plasmodium chabaudi</i> chabaudi	SAR Group	341
<i>Plasmodium falciparum</i> 3D7	SAR Group	336
<i>Emiliana huxleyi</i> CCMP1516	Excavata Plus	529
<i>Guillardia theta</i> CCMP2712	Excavata Plus	549
<i>Trichomonas vaginalis</i> G3	Excavata Plus	175
<i>Tritrichomonas foetus</i>	Excavata Plus	175
<i>Giardia lamblia</i> ATCC 50803	Excavata Plus	116
<i>Naegleria gruberi</i> strain NEG-M	Excavata Plus	495
<i>Angomonas deanei</i>	Excavata Plus	398
<i>Trypanosoma cruzi</i> strain CL Brener	Excavata Plus	395
<i>Trypanosoma brucei</i> brucei TREU927	Excavata Plus	375
<i>Leishmania major</i> strain Friedlin	Excavata Plus	402
<i>Leishmania infantum</i> JPCM5	Excavata Plus	397
<i>Monocercomonoides</i>	Excavata Plus	114
<i>Archaeoglobus fulgidus</i> DSM 4304	Archaea	93
<i>Ferroplasma acidophilum</i> DSM 10642	Archaea	91
<i>Geoglobus ahangari</i>	Archaea	90
<i>Hadesarchaea archaeon</i> YNP_N21	Archaea	60
<i>Halobacterium salinarum</i> R1	Archaea	124
<i>Natronomonas pharaonis</i> DSM 2160	Archaea	130
<i>Haloarcula marismortui</i> ATCC 43049	Archaea	129
<i>Natrinema pellirubrum</i> DSM 15624	Archaea	129
<i>Methanobrevibacter smithii</i> ATCC 35061	Archaea	71
<i>Methanothermobacter thermautotrophicus</i> str. Delta H	Archaea	76
<i>Methanosphaera stadtmanae</i> DSM 3091	Archaea	70
<i>Methanocaldococcus jannaschii</i> DSM 2661	Archaea	64
<i>Methanococcus maripaludis</i> C5	Archaea	67
<i>Methanosarcina acetivorans</i> C2A	Archaea	104
<i>Methanoregula boonei</i> 6A8	Archaea	91
<i>Methanoplanus limicola</i> DSM 2279	Archaea	87

Species	Group	No. orthologues
<i>Methanopyrus kandleri</i> AV19	Archaea	61
<i>Pyrococcus furiosus</i> DSM 3638	Archaea	81
<i>Thermococcus kodakarensis</i> KOD1	Archaea	80
<i>Palaeococcus pacificus</i> DY20341	Archaea	87
<i>Thermoplasma volcanium</i> GSS1	Archaea	103
<i>Ferroplasma acidarmanus</i> fer1	Archaea	110
<i>Picrophilus torridus</i> DSM 9790	Archaea	109
<i>Aeropyrum pernix</i> K1	Archaea	102
<i>Pyrobaculum aerophilum</i> str. IM2	Archaea	97
<i>Sulfolobus tokodaii</i> str. 7	Archaea	107
<i>Ignicoccus hospitalis</i>	Archaea	69
<i>Staphylothermus marinus</i> F1	Archaea	68
<i>Cenarchaeum symbiosum</i> A	Archaea	93
<i>Nitrosopumilus maritimus</i> SCM1	Archaea	100
<i>Nitrososphaera viennensis</i> EN76	Archaea	101
<i>Candidatus Korarchaeum cryptofilum</i> OPF8	Archaea	85
<i>Candidatus Haloredivivus</i> sp. G17	Archaea	49
<i>Candidatus Nanosalinarum</i> sp. J07AB56	Archaea	54
<i>Candidatus Parvarchaeum acidophilus</i> ARMAN-5	Archaea	61
<i>Nanoarchaeum equitans</i> Kin4-M	Archaea	40
<i>Lokiarchaeum</i> sp. GC14_75	Archaea	105
<i>Planctomyces maris</i> DSM 8797	Bacteria	203
<i>Isosphaera pallida</i> ATCC 43644	Bacteria	229
<i>Rhodopirellula baltica</i> SH 1	Bacteria	209
<i>Chlamydomyces pneumoniae</i> CWL029	Bacteria	143
<i>Chlamydia trachomatis</i>	Bacteria	137
<i>Waddlia chondrophila</i> WSU 86-1044	Bacteria	196
<i>Fervidobacterium nodosum</i> Rt17-B1	Bacteria	171
<i>Kosmotoga olearia</i> TBF 19.5.1	Bacteria	170
<i>Petrotoga mobilis</i> SJ95	Bacteria	174
<i>Thermosiphon africanus</i> TCF52B	Bacteria	169
<i>Thermotoga maritima</i> MSB8	Bacteria	160
<i>Mesotoga prima</i> MesG1.Ag.4.2	Bacteria	176
<i>Dictyoglomus thermophilum</i> H-6-12	Bacteria	159
<i>Dictyoglomus turgidum</i> DSM 6724	Bacteria	154
<i>Aquifex aeolicus</i> VF5	Bacteria	189
<i>Persephonella marina</i> EX-H1	Bacteria	191
<i>Thermovibrio ammonificans</i> HB-1	Bacteria	176
<i>Leptospirillum rubrum</i>	Bacteria	190
<i>Leptospirillum ferro-diazotrophum</i>	Bacteria	194
<i>Thermodesulfobacterium yellowstonii</i> DSM 11347	Bacteria	179
<i>Mycoplasma genitalium</i> G37	Bacteria	104

Species	Group	No. orthologues
<i>Onion yellows phytoplasma OY-M</i>	Bacteria	93
<i>Sealdella termitidis</i> ATCC 33386	Bacteria	182
<i>Leptotrichia buccalis</i> C-1013-b	Bacteria	163
<i>Fusobacterium nucleatum</i> subsp. <i>nucleatum</i> ATCC 25586	Bacteria	160
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	Bacteria	223
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> Mu50	Bacteria	207
<i>Listeria monocytogenes</i> EGD-e	Bacteria	196
<i>Streptococcus agalactiae</i>	Bacteria	175
<i>Clostridium tetani</i> E88	Bacteria	163
<i>Clostridium acetobutylicum</i> ATCC 824	Bacteria	187
<i>Thermoanaerobacter tengcongensis</i> MB4	Bacteria	193
<i>Mycobacterium tuberculosis</i> CDC1551	Bacteria	227
<i>Corynebacterium efficiens</i> YS-314	Bacteria	202
<i>Streptomyces coelicolor</i>	Bacteria	249
<i>Thermosynechococcus elongatus</i> BP-1	Bacteria	209
<i>Gloeobacter violaceus</i> PCC 7421	Bacteria	225
<i>Prochlorococcus marinus</i> str. MIT 9313	Bacteria	193
<i>Roseiflexus castenholzii</i> DSM 13941	Bacteria	239
<i>Chloroflexus aggregans</i> DSM 9485	Bacteria	235
<i>Herpetosiphon aurantiacus</i> DSM 785	Bacteria	241
<i>Anaerolinea thermophila</i> UNI-1	Bacteria	205
<i>Ktedonobacter racemifer</i> DSM 44963	Bacteria	240
<i>Sphaerobacter thermophilus</i> DSM 20745	Bacteria	228
<i>Thermus thermophilus</i> HB8	Bacteria	222
<i>Meiothermus silvanus</i> DSM 9946	Bacteria	227
<i>Deinococcus radiodurans</i> R1	Bacteria	222
<i>Methylophilum infernorum</i> V4	Bacteria	197
<i>Opitutus terrae</i> PB90-1	Bacteria	222
<i>Akkermansia muciniphila</i> ATCC BAA-835	Bacteria	180
<i>Porphyromonas gingivalis</i> W83	Bacteria	163
<i>Bacteroides thetaiotaomicron</i> VPI-5482	Bacteria	198
<i>Flavobacterium psychrophilum</i>	Bacteria	216
<i>Chloroherpeton thalassium</i> ATCC 35110	Bacteria	221
<i>Chlorobium tepidum</i> TLS	Bacteria	194
<i>Prosthecochloris aestuarii</i> DSM 271	Bacteria	203
<i>Treponema denticola</i> ATCC 35405	Bacteria	151
<i>Borrelia burgdorferi</i> B31	Bacteria	122
<i>Brachyspira murdochii</i> DSM 12563	Bacteria	178
<i>Leptospira interrogans</i> serovar Lai str. 56601	Bacteria	217
<i>Acidobacterium capsulatum</i> ATCC 51196	Bacteria	221
<i>Candidatus Koribacter versatilis</i> Ellin345	Bacteria	229
<i>Terriglobus saanensis</i> SP1PR4	Bacteria	214

Species	Group	No. orthologues
<i>Gemmatimonas aurantiaca</i> T-27	Bacteria	239
<i>Agrobacterium tumefaciens</i> str. C58	Bacteria	280
<i>Rhizobium etli</i> CFN 42	Bacteria	292
<i>Rickettsia prowazekii</i> str. Madrid E	Bacteria	198
<i>Nitrosospira multififormis</i> ATCC 25196	Bacteria	232
<i>Nitrosomonas europaea</i> ATCC 19718	Bacteria	233
<i>Burkholderia mallei</i> ATCC 23344	Bacteria	260
<i>Neisseria meningitidis</i> ATCC 13091	Bacteria	212
<i>Pseudomonas syringae</i> pv. tomato str. DC3000	Bacteria	265
<i>Pseudomonas aeruginosa</i> PA7	Bacteria	290
<i>Escherichia coli</i> BW25113	Bacteria	251
<i>Vibrio cholerae</i> O1 biovar El Tor str. N16961	Bacteria	224
<i>Bdellovibrio bacteriovorus</i> HD100	Bacteria	231
<i>Geobacter sulfurreducens</i> PCA	Bacteria	223
<i>Lawsonia intracellularis</i> N343	Bacteria	146
<i>Helicobacter winthamensis</i> ATCC BAA-430	Bacteria	174
<i>Helicobacter pylori</i> J99	Bacteria	162
<i>Wolinella succinogenes</i> DSM 1740	Bacteria	183
<i>Campylobacter jejuni</i> subsp. jejuni NCTC 11168	Bacteria	176
<i>Aminobacterium colombiense</i> DSM 12261	Bacteria	181
<i>Anaerobaculum hydrogeniformans</i> ATCC BAA-1850	Bacteria	178
<i>Dethiosulfovibrio peptidovorans</i> DSM 11002	Bacteria	183
<i>Calditerrivibrio nitroreducens</i> DSM 19672	Bacteria	201
<i>Deferribacter desulfuricans</i> SSM1	Bacteria	210
<i>Denitrovibrio acetiphilus</i> DSM 12809	Bacteria	214
<i>Thermodesulfatator indicus</i> DSM 15286	Bacteria	172
<i>Thermodesulfobacterium commune</i> DSM 2178	Bacteria	168
<i>Desulfurispirillum indicum</i> S5	Bacteria	210

Table 2. Numbers of predicted orthologues of each of the 1,550 MPI 2017 genes in the manually curated dataset. Numbers of predicted orthologues are also shown for eight phylogenetic groups: Hz = Holozoa, Hm = Holomycota, Ab = Amoebozoa, Ap = Archaeplastida, SG = SAR group, Ep = Excavata Plus, A = Archaea, B = Bacteria

EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	SG	Ep	A	B	EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	SG	Ep	A	B
ENSG00000109576	<i>AADAT</i>	210	80	35	6	9	15	2	15	48	ENSG00000131473	<i>ACLY</i>	180	92	31	6	30	16	0	0	5
ENSG00000205002	<i>AARD</i>	21	21	0	0	0	0	0	0	0	ENSG00000100412	<i>ACO2</i>	138	95	37	6	0	0	0	0	0
ENSG00000124608	<i>AARS2</i>	38	38	0	0	0	0	0	0	0	ENSG00000184227	<i>ACOT1</i>	1	1	0	0	0	0	0	0	0
ENSG000000008311	<i>AASS</i>	140	84	30	6	0	18	2	0	0	ENSG00000112304	<i>ACOT13</i>	194	81	28	5	26	12	6	11	25
ENSG00000183044	<i>ABAT</i>	135	84	26	6	6	11	2	0	0	ENSG00000119673	<i>ACOT2</i>	43	43	0	0	0	0	0	0	0
ENSG00000179869	<i>ABCA13</i>	27	27	0	0	0	0	0	0	0	ENSG00000097021	<i>ACOT7</i>	57	56	1	0	0	0	0	0	0
ENSG00000141338	<i>ABCA8</i>	28	28	0	0	0	0	0	0	0	ENSG00000101473	<i>ACOT8</i>	137	66	37	0	21	1	2	0	10
ENSG00000154258	<i>ABCA9</i>	16	16	0	0	0	0	0	0	0	ENSG00000123130	<i>ACOT9</i>	158	74	28	6	31	17	2	0	0
ENSG00000135776	<i>ABCB10</i>	283	80	33	4	31	25	10	28	72	ENSG00000162836	<i>ACP6</i>	136	64	26	6	11	20	9	0	0
ENSG00000115657	<i>ABCB6</i>	138	81	30	6	6	13	2	0	0	ENSG00000167107	<i>ACSF2</i>	85	80	4	0	1	0	0	0	0
ENSG00000131269	<i>ABCB7</i>	212	96	42	6	32	29	7	0	0	ENSG00000176715	<i>ACSF3</i>	240	80	25	6	31	19	8	19	52
ENSG00000197150	<i>ABCB8</i>	87	79	4	4	0	0	0	0	0	ENSG00000151726	<i>ACSL1</i>	234	89	40	3	30	36	12	4	20
ENSG00000140798	<i>ABCC12</i>	39	39	0	0	0	0	0	0	0	ENSG00000123983	<i>ACSL3</i>	43	43	0	0	0	0	0	0	0
ENSG00000101986	<i>ABCD1</i>	34	34	0	0	0	0	0	0	0	ENSG00000068366	<i>ACSL4</i>	162	89	36	7	23	7	0	0	0
ENSG00000117528	<i>ABCD3</i>	160	81	5	6	32	26	10	0	0	ENSG00000197142	<i>ACSL5</i>	41	41	0	0	0	0	0	0	0
ENSG00000164163	<i>ABCE1</i>	253	90	42	8	31	35	11	36	0	ENSG00000164398	<i>ACSL6</i>	43	43	0	0	0	0	0	0	0
ENSG00000118777	<i>ABCF2</i>	183	87	36	6	31	14	9	0	0	ENSG00000166743	<i>ACSM1</i>	21	21	0	0	0	0	0	0	0
ENSG00000118777	<i>ABCG2</i>	199	83	40	8	32	28	8	0	0	ENSG00000183747	<i>ACSM2A</i>	23	23	0	0	0	0	0	0	0
ENSG00000144827	<i>ABHD10</i>	56	51	5	0	0	0	0	0	0	ENSG00000066813	<i>ACSM2B</i>	2	2	0	0	0	0	0	0	0
ENSG00000106077	<i>ABHD11</i>	226	94	36	6	31	23	8	0	28	ENSG00000005187	<i>ACSM3</i>	91	56	0	0	0	0	0	17	18
ENSG00000204427	<i>ABHD16A</i>	90	90	0	0	0	0	0	0	0	ENSG00000215009	<i>ACSM4</i>	30	30	0	0	0	0	0	0	0
ENSG00000060971	<i>ACA41</i>	197	57	38	6	33	25	2	7	29	ENSG00000183549	<i>ACSM5</i>	23	23	0	0	0	0	0	0	0
ENSG00000167315	<i>ACA42</i>	115	86	18	0	0	11	0	0	0	ENSG00000173124	<i>ACSM6</i>	10	10	0	0	0	0	0	0	0
ENSG00000076555	<i>ACACB</i>	39	39	0	0	0	0	0	0	0	ENSG00000154930	<i>ACSSI</i>	134	65	43	5	0	19	2	0	0
ENSG00000111271	<i>ACAD10</i>	92	58	29	1	0	4	0	0	0	ENSG00000111058	<i>ACSS3</i>	116	75	14	6	0	13	1	0	7
ENSG00000240303	<i>ACAD11</i>	54	54	0	0	0	0	0	0	0	ENSG00000075624	<i>ACTB</i>	217	91	41	8	28	36	12	1	0
ENSG00000151498	<i>ACAD8</i>	90	67	6	6	6	5	0	0	0	ENSG00000072110	<i>ACTN1</i>	145	92	36	5	0	8	4	0	0
ENSG00000177646	<i>ACAD9</i>	63	57	0	0	0	6	0	0	0	ENSG00000170634	<i>ACTP2</i>	203	80	13	1	30	19	9	0	51
ENSG00000115361	<i>ACADL</i>	63	63	0	0	0	0	0	0	0	ENSG00000063761	<i>ADCK1</i>	255	85	35	5	33	23	6	20	48
ENSG00000117054	<i>ACADM</i>	112	80	18	0	0	9	5	0	0	ENSG00000133597	<i>ADCK2</i>	160	60	38	6	29	19	8	0	0
ENSG00000122971	<i>ACADS</i>	89	75	0	0	0	12	2	0	0	ENSG00000163050	<i>ADCK3</i>	200	92	35	6	31	29	7	0	0
ENSG00000196177	<i>ACADSB</i>	213	80	29	6	3	16	7	16	56	ENSG00000123815	<i>ADCK4</i>	28	28	0	0	0	0	0	0	0
ENSG00000072778	<i>ACADVL</i>	89	88	1	0	0	0	0	0	0	ENSG00000173137	<i>ADCK5</i>	97	74	9	3	0	4	7	0	0
ENSG00000075239	<i>ACAT1</i>	194	90	37	6	28	26	7	0	0	ENSG00000143199	<i>ADCY10</i>	100	57	4	5	0	31	3	0	0
ENSG00000182827	<i>ACBD3</i>	93	93	0	0	0	0	0	0	0	ENSG00000118492	<i>ADGB</i>	71	71	0	0	0	0	0	0	0

EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000117114	ADGRL2	44	44	0	0	0	0	0	0	0	211	91	42	8	33	30	7	0	0
ENSG00000172955	ADH6	12	12	0	0	0	0	0	0	0	277	87	37	6	32	30	9	15	61
ENSG00000147576	ADHFE1	112	74	24	6	0	5	0	2	1	106	82	4	0	10	9	1	0	0
ENSG00000181915	ADO	116	84	2	4	0	19	7	0	0	43	43	0	0	0	0	0	0	0
ENSG00000116863	ADPRHL2	129	80	4	1	5	7	1	9	22	35	35	0	0	0	0	0	0	0
ENSG00000141385	AFG3L2	304	96	37	6	32	34	9	0	90	44	44	0	0	0	0	0	0	0
ENSG00000006530	AGK	86	86	0	0	0	0	0	0	0	44	44	0	0	0	0	0	0	0
ENSG00000116771	AGMAT	157	57	25	0	1	7	0	37	30	215	91	37	8	32	35	12	0	0
ENSG0000026652	AGPAT4	78	78	0	0	0	0	0	0	0	192	94	33	7	28	20	10	0	0
ENSG00000155189	AGPAT5	78	78	0	0	0	0	0	0	0	190	89	9	7	33	30	11	1	10
ENSG00000172482	AGXT	242	70	33	6	32	6	1	31	63	143	72	32	0	32	7	0	0	0
ENSG00000113492	AGXT2	109	78	0	1	25	4	1	0	0	277	91	37	8	7	21	5	32	76
ENSG00000156709	AIFM1	98	95	3	0	0	0	0	0	0	44	44	0	0	0	0	0	0	0
ENSG00000183773	AIFM3	140	76	28	6	3	22	2	0	3	68	68	0	0	0	0	0	0	0
ENSG00000004455	AK2	305	88	40	6	33	36	11	14	77	89	76	13	0	0	0	0	0	0
ENSG00000147853	AK3	104	78	21	0	0	5	0	0	0	102	72	30	0	0	0	0	0	0
ENSG00000162433	AK4	43	43	0	0	0	0	0	0	0	38	38	0	0	0	0	0	0	0
ENSG00000121057	AKAP1	78	78	0	0	0	0	0	0	0	171	85	34	2	2	20	8	2	18
ENSG00000108599	AKAP10	86	86	0	0	0	0	0	0	0	97	76	0	0	21	0	0	0	0
ENSG00000085662	AKR1B1	26	26	0	0	0	0	0	0	0	201	90	30	6	32	32	11	0	0
ENSG00000227471	AKR1B15	2	2	0	0	0	0	0	0	0	99	76	3	0	5	11	4	0	0
ENSG00000198610	AKR1C4	8	8	0	0	0	0	0	0	0	61	61	0	0	0	0	0	0	0
ENSG0000053371	AKR7A2	7	7	0	0	0	0	0	0	0	55	55	0	0	0	0	0	0	0
ENSG00000142208	AKT1	36	36	0	0	0	0	0	0	0	22	22	0	0	0	0	0	0	0
ENSG0000023330	ALAS1	42	42	0	0	0	0	0	0	0	24	24	0	0	0	0	0	0	0
ENSG00000158578	ALAS2	142	81	35	5	0	18	0	0	3	24	24	0	0	0	0	0	0	0
ENSG0000059573	ALDH18A1	249	85	34	1	30	21	6	0	72	9	9	0	0	0	0	0	0	0
ENSG00000137124	ALDH1B1	27	27	0	0	0	0	0	0	0	82	70	2	5	0	4	1	0	0
ENSG00000144908	ALDH1L1	23	23	0	0	0	0	0	0	0	143	83	18	5	26	9	1	0	1
ENSG00000136010	ALDH1L2	76	75	0	0	1	0	0	0	0	39	39	0	0	0	0	0	0	0
ENSG0000011275	ALDH2	187	77	30	6	25	13	6	0	30	52	51	0	0	0	1	0	0	0
ENSG0000072210	ALDH3A2	188	92	32	8	28	19	9	0	0	266	77	33	0	33	14	5	30	74
ENSG00000159423	ALDH4A1	150	88	36	6	3	1	6	0	10	184	89	37	5	32	14	7	0	0
ENSG00000112294	ALDH5A1	203	92	32	0	30	20	7	0	22	156	93	1	0	29	33	0	0	0
ENSG00000119711	ALDH6A1	198	87	33	6	27	23	3	3	16	2	2	0	0	0	0	0	0	0
ENSG00000164904	ALDH7A1	155	83	17	6	23	14	1	2	9	45	45	0	0	0	0	0	0	0
ENSG00000143149	ALDH9A1	72	72	0	0	0	0	0	0	0	167	92	33	6	20	14	2	0	0
ENSG00000100601	ALKBH1	180	88	27	5	28	24	8	0	0	74	74	0	0	0	0	0	0	0
ENSG00000166199	ALKBH3	96	59	24	0	0	10	3	0	0	264	81	35	6	32	16	2	8	84
ENSG00000125652	ALKBH7	87	82	0	0	0	5	0	0	0	280	97	36	1	30	32	6	0	78
ENSG00000242110	AMACR	114	81	21	4	0	8	0	0	0	288	96	34	6	33	34	7	0	78
ENSG00000110497	AMBRA1	88	76	0	0	0	11	1	0	0	290	96	37	7	33	32	7	0	78

EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B	EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000099624	<i>ATP5D</i>	242	90	36	6	30	23	8	0	49	ENSG00000029571	<i>BLID</i>	11	11	0	0	0	0	0	0	0
ENSG00000124172	<i>ATP5E</i>	146	69	24	6	31	10	6	0	0	ENSG00000135441	<i>BLOC1S1</i>	148	88	12	8	29	6	5	0	0
ENSG00000180389	<i>ATP5EP2</i>	2	0	0	0	0	0	0	0	0	ENSG000000196072	<i>BLOC1S2</i>	133	85	8	8	25	2	5	0	0
ENSG00000116459	<i>ATP5FI</i>	123	96	26	0	1	0	0	0	0	ENSG00000113734	<i>BNIP1</i>	96	89	0	6	0	0	1	0	0
ENSG00000159199	<i>ATP5G1</i>	31	31	0	0	0	0	0	0	0	ENSG00000176171	<i>BNIP3</i>	67	67	0	0	0	0	0	0	0
ENSG00000135390	<i>ATP5G2</i>	194	88	19	2	27	22	5	0	31	ENSG00000104765	<i>BNIP3L</i>	43	43	0	0	0	0	0	0	0
ENSG00000154518	<i>ATP5G3</i>	25	25	0	0	0	0	0	0	0	ENSG00000178096	<i>BOLA1</i>	205	82	38	6	32	22	6	4	15
ENSG00000167863	<i>ATP5H</i>	141	88	30	0	20	3	0	0	0	ENSG00000163170	<i>BOLA3</i>	122	82	31	6	0	3	0	0	0
ENSG00000169020	<i>ATP5I</i>	69	68	1	0	0	0	0	0	0	ENSG00000137274	<i>BPFL</i>	82	77	0	0	0	4	1	0	0
ENSG00000154723	<i>ATP5J</i>	84	84	0	0	0	0	0	0	0	ENSG00000162813	<i>BPNT1</i>	130	94	1	7	0	19	9	0	0
ENSG00000241468	<i>ATP5L2</i>	77	77	0	0	0	0	0	0	0	ENSG00000184992	<i>BR3BP</i>	42	42	0	0	0	0	0	0	0
ENSG00000167283	<i>ATP5L</i>	133	91	21	1	20	0	0	0	0	ENSG00000162670	<i>BRINP3</i>	43	43	0	0	0	0	0	0	0
ENSG00000249222	<i>ATP5L2</i>	4	4	0	0	0	0	0	0	0	ENSG00000172270	<i>BSG</i>	40	40	0	0	0	0	0	0	0
ENSG00000241837	<i>ATP5O</i>	255	95	34	4	33	31	2	0	56	ENSG00000138152	<i>BTBD16</i>	46	46	0	0	0	0	0	0	0
ENSG00000125375	<i>ATP5S</i>	84	82	2	0	0	0	0	0	0	ENSG00000005379	<i>BZRAP1</i>	39	39	0	0	0	0	0	0	0
ENSG00000105341	<i>ATP5SL</i>	70	70	0	0	0	0	0	0	0	ENSG00000165507	<i>C10orf10</i>	33	33	0	0	0	0	0	0	0
ENSG00000114573	<i>ATP6V1A</i>	277	97	45	8	33	35	11	28	20	ENSG00000107815	<i>C10orf2</i>	129	87	6	5	3	26	2	0	0
ENSG00000123191	<i>ATP7B</i>	130	34	0	0	0	0	0	25	71	ENSG00000139637	<i>C12orf10</i>	214	89	42	7	33	31	12	0	0
ENSG00000123472	<i>ATPAF1</i>	162	73	32	6	26	18	7	0	0	ENSG00000130921	<i>C12orf65</i>	132	83	29	1	3	15	1	0	0
ENSG00000171953	<i>ATPAF2</i>	202	96	37	6	32	21	8	0	2	ENSG00000179933	<i>C14orf119</i>	61	61	0	0	0	0	0	0	0
ENSG00000130770	<i>ATPFI</i>	98	85	11	0	0	1	1	0	0	ENSG00000135943	<i>C14orf159</i>	52	52	0	0	0	0	0	0	0
ENSG00000148090	<i>AUHL</i>	158	79	28	5	26	16	4	0	0	ENSG00000156411	<i>C14orf2</i>	38	38	0	0	0	0	0	0	0
ENSG00000175756	<i>AURKAIPI</i>	42	42	0	0	0	0	0	0	0	ENSG00000189227	<i>C15orf61</i>	65	65	0	0	0	0	0	0	0
ENSG00000002330	<i>BAD</i>	37	37	0	0	0	0	0	0	0	ENSG00000188277	<i>C15orf62</i>	32	32	0	0	0	0	0	0	0
ENSG00000112208	<i>BAG2</i>	70	70	0	0	0	0	0	0	0	ENSG00000174109	<i>C16orf91</i>	49	49	0	0	0	0	0	0	0
ENSG00000166170	<i>BAG5</i>	44	44	0	0	0	0	0	0	0	ENSG00000178927	<i>C17orf62</i>	55	55	0	0	0	0	0	0	0
ENSG00000007516	<i>BALAP3</i>	69	69	0	0	0	0	0	0	0	ENSG000000224877	<i>C17orf89</i>	28	28	0	0	0	0	0	0	0
ENSG00000030110	<i>BAKI</i>	38	38	0	0	0	0	0	0	0	ENSG00000131943	<i>C19orf12</i>	57	57	0	0	0	0	0	0	0
ENSG00000087088	<i>BAX</i>	66	66	0	0	0	0	0	0	0	ENSG00000142444	<i>C19orf52</i>	89	89	0	0	0	0	0	0	0
ENSG00000105327	<i>BBC3</i>	20	20	0	0	0	0	0	0	0	ENSG00000174917	<i>C19orf70</i>	61	61	0	0	0	0	0	0	0
ENSG00000129151	<i>BBOX1</i>	82	75	5	1	0	1	0	0	0	ENSG00000162398	<i>C1orf177</i>	64	64	0	0	0	0	0	0	0
ENSG00000060982	<i>BCAT1</i>	290	91	37	6	33	24	11	19	69	ENSG000000203724	<i>C1orf53</i>	60	57	3	0	0	0	0	0	0
ENSG00000105552	<i>BCAT2</i>	23	23	0	0	0	0	0	0	0	ENSG00000108561	<i>C1QBP</i>	140	87	21	0	16	11	5	0	0
ENSG00000248098	<i>BCKDHA</i>	196	88	29	5	27	30	7	10	0	ENSG00000088854	<i>C20orf194</i>	69	62	0	0	0	6	1	0	0
ENSG00000083123	<i>BCKDHB</i>	196	89	27	6	27	30	7	10	0	ENSG00000160221	<i>C21orf33</i>	82	69	0	0	0	0	0	0	13
ENSG00000103507	<i>BCKDK</i>	102	68	24	0	3	6	1	0	0	ENSG000000215012	<i>C22orf29</i>	16	16	0	0	0	0	0	0	0
ENSG00000171791	<i>BCL2</i>	41	41	0	0	0	0	0	0	0	ENSG00000162972	<i>C2orf47</i>	70	70	0	0	0	0	0	0	0
ENSG00000171552	<i>BCL2L1</i>	66	66	0	0	0	0	0	0	0	ENSG00000178074	<i>C2orf69</i>	108	79	0	0	16	11	2	0	0
ENSG00000099968	<i>BCL2L13</i>	44	44	0	0	0	0	0	0	0	ENSG00000131379	<i>C3orf20</i>	49	49	0	0	0	0	0	0	0
ENSG00000197580	<i>BCO2</i>	42	42	0	0	0	0	0	0	0	ENSG00000174928	<i>C3orf33</i>	72	72	0	0	0	0	0	0	0
ENSG00000074582	<i>BCSL1</i>	154	94	35	6	0	11	8	0	0	ENSG00000164241	<i>C5orf63</i>	56	56	0	0	0	0	0	0	0
ENSG00000161267	<i>BDHI</i>	64	64	0	0	0	0	0	0	0	ENSG000000204564	<i>C6orf136</i>	75	74	1	0	0	0	0	0	0

EnsemblID	Gene	Sum	Hx	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000130349	<i>C6orf203</i>	67	67	0	0	0	0	0	0	0
ENSG00000164898	<i>C7orf55</i>	69	69	0	0	0	0	0	0	0
ENSG00000164645	<i>C7orf62</i>	62	58	4	0	0	0	0	0	0
ENSG00000213563	<i>C8orf82</i>	99	75	9	0	0	11	4	0	0
ENSG00000165233	<i>C9orf89</i>	40	40	0	0	0	0	0	0	0
ENSG00000174990	<i>C45A</i>	25	25	0	0	0	0	0	0	0
ENSG00000169239	<i>C45B</i>	42	42	0	0	0	0	0	0	0
ENSG00000179218	<i>CALR</i>	164	94	8	8	27	17	10	0	0
ENSG00000127022	<i>CANX</i>	175	92	30	4	29	14	6	0	0
ENSG00000142330	<i>CAPN10</i>	43	43	0	0	0	0	0	0	0
ENSG00000110888	<i>CAPRN2</i>	43	43	0	0	0	0	0	0	0
ENSG00000213995	<i>CARKD</i>	328	90	43	8	33	30	8	34	82
ENSG00000134905	<i>CARS2</i>	87	72	1	3	11	0	0	0	0
ENSG00000145439	<i>CBR4</i>	227	51	6	0	33	19	8	22	88
ENSG00000136682	<i>CBWD2</i>	1	1	0	0	0	0	0	0	0
ENSG00000171097	<i>CCBL1</i>	42	42	0	0	0	0	0	0	0
ENSG00000137944	<i>CCBL2</i>	271	90	35	6	2	20	7	31	80
ENSG0000005059	<i>CCDC109B</i>	41	41	0	0	0	0	0	0	0
ENSG00000164366	<i>CCDC127</i>	43	43	0	0	0	0	0	0	0
ENSG00000164051	<i>CCDC51</i>	84	84	0	0	0	0	0	0	0
ENSG00000160124	<i>CCDC58</i>	99	84	15	0	0	0	0	0	0
ENSG00000137500	<i>CCDC90B</i>	43	43	0	0	0	0	0	0	0
ENSG00000173992	<i>CCS</i>	93	68	11	0	14	0	0	0	0
ENSG0000004468	<i>CD38</i>	48	48	0	0	0	0	0	0	0
ENSG00000158402	<i>CD3C</i>	32	32	0	0	0	0	0	0	0
ENSG00000103502	<i>CDIPT</i>	216	90	44	7	33	31	11	0	0
ENSG00000170312	<i>CDKI</i>	86	86	0	0	0	0	0	0	0
ENSG00000101391	<i>CDK5RAP1</i>	242	89	9	6	28	20	4	0	86
ENSG00000218739	<i>CEBPZOS</i>	9	9	0	0	0	0	0	0	0
ENSG00000069998	<i>CECR5</i>	170	78	32	6	27	19	8	0	0
ENSG00000184524	<i>CEND1</i>	30	30	0	0	0	0	0	0	0
ENSG00000121289	<i>CEP89</i>	69	69	0	0	0	0	0	0	0
ENSG00000100422	<i>CERK</i>	108	85	0	0	23	0	0	0	0
ENSG00000172586	<i>CHCHD1</i>	82	78	0	1	3	0	0	0	0
ENSG00000250479	<i>CHCHD10</i>	34	34	0	0	0	0	0	0	0
ENSG00000106153	<i>CHCHD2</i>	130	84	27	3	6	9	1	0	0
ENSG00000106554	<i>CHCHD3</i>	80	80	0	0	0	0	0	0	0
ENSG00000163528	<i>CHCHD4</i>	148	82	34	4	28	0	0	0	0
ENSG00000125611	<i>CHCHD5</i>	44	44	0	0	0	0	0	0	0
ENSG00000159685	<i>CHCHD6</i>	42	42	0	0	0	0	0	0	0
ENSG00000170791	<i>CHCHD7</i>	92	69	23	0	0	0	0	0	0
ENSG00000016391	<i>CHDH</i>	158	86	26	6	9	16	8	0	7
ENSG000000122873	<i>CISD1</i>	44	44	0	0	0	0	0	0	0
ENSG00000145354	<i>CISD2</i>	91	91	0	0	0	0	0	0	0
ENSG00000277972	<i>CISD3</i>	129	84	6	5	0	29	5	0	0
ENSG00000223572	<i>CKMT1A</i>	32	32	0	0	0	0	0	0	0
ENSG00000237289	<i>CKMT1B</i>	33	33	0	0	0	0	0	0	0
ENSG00000131730	<i>CKMT2</i>	49	49	0	0	0	0	0	0	0
ENSG00000213719	<i>CLIC1</i>	38	38	0	0	0	0	0	0	0
ENSG00000169504	<i>CLIC4</i>	41	41	0	0	0	0	0	0	0
ENSG00000112782	<i>CLIC5</i>	84	84	0	0	0	0	0	0	0
ENSG00000162129	<i>CLPB</i>	267	85	44	8	32	4	4	0	90
ENSG00000125656	<i>CLPP</i>	259	87	26	0	33	22	2	0	89
ENSG00000166855	<i>CLPX</i>	281	95	26	5	32	27	8	0	88
ENSG00000141367	<i>CLTC</i>	220	95	37	8	33	36	11	0	0
ENSG00000132361	<i>CLUH</i>	175	91	31	6	20	22	5	0	0
ENSG00000125246	<i>CLYBL</i>	166	83	22	6	2	10	9	9	25
ENSG00000187118	<i>CMC1</i>	128	83	11	3	25	5	1	0	0
ENSG00000103121	<i>CMC2</i>	117	56	26	1	26	8	0	0	0
ENSG00000182712	<i>CMC4</i>	89	51	17	0	20	1	0	0	0
ENSG00000134326	<i>CMFK2</i>	56	51	0	0	0	0	0	0	0
ENSG00000173786	<i>CNP</i>	60	60	0	0	0	0	0	0	0
ENSG00000106603	<i>COA1</i>	62	60	1	0	0	0	0	0	1
ENSG00000183978	<i>COA3</i>	61	50	11	0	0	0	0	0	0
ENSG00000181924	<i>COA4</i>	112	73	18	0	20	1	0	0	0
ENSG00000183513	<i>COA5</i>	151	79	28	4	25	14	1	0	0
ENSG00000168275	<i>COA6</i>	127	76	30	6	14	1	0	0	0
ENSG00000162377	<i>COA7</i>	100	89	0	0	0	11	0	0	0
ENSG00000068120	<i>COASY</i>	182	90	25	8	30	20	9	0	0
ENSG00000093010	<i>COMT</i>	98	48	13	1	9	9	2	0	16
ENSG00000165644	<i>COMTD1</i>	175	67	20	6	32	7	4	7	32
ENSG00000135469	<i>COQ10A</i>	35	35	0	0	0	0	0	0	0
ENSG00000115520	<i>COQ10B</i>	182	88	31	6	31	20	6	0	0
ENSG00000173085	<i>COQ2</i>	215	91	34	4	32	30	8	0	16
ENSG00000132423	<i>COQ3</i>	217	93	34	6	32	33	8	0	11
ENSG00000167113	<i>COQ4</i>	198	88	34	6	33	27	8	0	2
ENSG00000110871	<i>COQ5</i>	274	91	35	6	33	30	8	1	70
ENSG00000119723	<i>COQ6</i>	184	90	35	5	33	15	6	0	0
ENSG00000167186	<i>COQ7</i>	150	89	30	5	5	11	7	0	3
ENSG00000086882	<i>COQ9</i>	170	81	26	6	33	15	8	0	1
ENSG00000167549	<i>CORO6</i>	152	84	33	8	0	18	9	0	0
ENSG0000006695	<i>COX10</i>	265	93	36	6	33	30	8	16	43
ENSG00000166260	<i>COX11</i>	213	95	36	6	32	29	7	0	8
ENSG00000178449	<i>COX14</i>	42	42	0	0	0	0	0	0	0

EnsemblID	Gene	Sum	Hr	Hm	Ab	Ap	Sg	Ep	A	B	EnsemblID	Gene	Sum	Hr	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000014919	COX15	217	96	37	6	29	31	8	1	9	ENSG000000159348	CYB5R1	33	33	0	0	0	0	0	0	0
ENSG000000133983	COX16	105	61	27	0	14	3	0	0	0	ENSG000000100243	CYB5R3	168	88	37	6	5	25	7	0	0
ENSG000000138495	COX17	172	76	32	6	31	20	7	0	0	ENSG000000179091	CYC1	207	92	36	6	32	30	8	0	3
ENSG000000163626	COX18	97	78	18	0	0	1	0	0	0	ENSG000000172115	CYCS	215	95	38	6	32	31	8	0	5
ENSG000000240230	COX19	174	71	35	6	30	24	8	0	0	ENSG000000140459	CYP11A1	42	42	0	0	0	0	0	0	0
ENSG000000203667	COX20	96	66	20	4	0	6	0	0	0	ENSG000000160882	CYP11B1	6	6	0	0	0	0	0	0	0
ENSG000000131143	COX41I	114	91	22	0	1	0	0	0	0	ENSG000000179142	CYP11B2	27	27	0	0	0	0	0	0	0
ENSG000000131055	COX412	31	31	0	0	0	0	0	0	0	ENSG000000140465	CYP11A1	41	41	0	0	0	0	0	0	0
ENSG000000178741	COX5A	128	93	35	0	0	0	0	0	0	ENSG000000019186	CYP24A1	70	70	0	0	0	0	0	0	0
ENSG000000135940	COX5B	181	92	34	1	31	21	2	0	0	ENSG000000135929	CYP27A1	43	43	0	0	0	0	0	0	0
ENSG000000111775	COX6A1	151	85	36	0	22	6	2	0	0	ENSG000000111012	CYP27B1	40	40	0	0	0	0	0	0	0
ENSG000000156885	COX6A2	24	24	0	0	0	0	0	0	0	ENSG000000180902	D2HGDH	184	85	35	6	31	21	6	0	0
ENSG000000126267	COX6B1	177	87	34	4	33	17	2	0	0	ENSG000000164488	DACT2	42	42	0	0	0	0	0	0	0
ENSG000000160471	COX6B2	19	19	0	0	0	0	0	0	0	ENSG000000182346	D4O4	9	9	0	0	0	0	0	0	0
ENSG000000164919	COX6C	63	63	0	0	0	0	0	0	0	ENSG000000132676	D4P3	186	92	30	5	29	23	7	0	0
ENSG000000161281	COX7A1	25	25	0	0	0	0	0	0	0	ENSG000000117593	D4RS2	258	89	31	6	33	6	1	1	91
ENSG000000112695	COX7A2	43	43	0	0	0	0	0	0	0	ENSG000000155368	DBI	172	77	30	5	33	21	2	0	4
ENSG000000115944	COX7A2L	72	66	6	0	0	0	0	0	0	ENSG000000137992	DBT	182	91	25	6	26	27	7	0	0
ENSG000000131174	COX7B	44	44	0	0	0	0	0	0	0	ENSG000000139990	DCAF5	84	72	12	0	0	0	0	0	0
ENSG000000170516	COX7B2	9	9	0	0	0	0	0	0	0	ENSG000000172992	DCAKD	281	82	41	1	32	27	10	0	88
ENSG000000127184	COX7C	83	66	17	0	0	0	0	0	0	ENSG000000169738	DCXR	78	75	0	0	0	3	0	0	0
ENSG000000176340	COX8A	35	35	0	0	0	0	0	0	0	ENSG000000168209	DDIT4	42	42	0	0	0	0	0	0	0
ENSG000000187581	COX8C	14	14	0	0	0	0	0	0	0	ENSG000000099977	DDIT4	48	48	0	0	0	0	0	0	0
ENSG000000080819	CPOX	208	87	36	6	33	23	5	0	18	ENSG000000182810	DDX28	72	72	0	0	0	0	0	0	0
ENSG000000021826	CPS1	201	51	0	0	33	25	7	18	67	ENSG000000104325	DECRI	71	71	0	0	0	0	0	0	0
ENSG000000110090	CPT1A	99	87	1	0	0	6	5	0	0	ENSG000000162777	DENND2D	42	42	0	0	0	0	0	0	0
ENSG000000205560	CPT1B	31	31	0	0	0	0	0	0	0	ENSG000000155792	DEPTOR	56	56	0	0	0	0	0	0	0
ENSG000000157184	CPT2	133	88	28	0	0	10	7	0	0	ENSG00000014956	DGUKOR	38	34	0	0	0	3	1	0	0
ENSG000000095321	CRAT	115	70	32	1	0	7	4	0	1	ENSG000000178700	DHERL1	3	3	0	0	0	0	0	0	0
ENSG000000118260	CREB1	43	43	0	0	0	0	0	0	0	ENSG000000102967	DHODH	231	91	32	1	30	32	3	4	38
ENSG000000088766	CRLS1	252	91	32	1	33	11	1	1	82	ENSG000000157379	DHRS1	80	68	0	1	2	8	1	0	0
ENSG000000008405	CRY1	196	76	24	0	32	14	1	9	40	ENSG000000100867	DHRS2	24	24	0	0	0	0	0	0	0
ENSG000000109846	CRY4B	40	40	0	0	0	0	0	0	0	ENSG000000162496	DHRS3	47	47	0	0	0	0	0	0	0
ENSG000000116791	CRYZ	77	71	0	1	0	1	1	1	2	ENSG000000157326	DHRS4	114	78	0	0	27	8	1	0	0
ENSG000000062485	CS	236	93	37	6	32	33	8	20	7	ENSG000000100612	DHRS7	134	70	2	6	29	24	3	0	0
ENSG000000047230	CTPS2	35	35	0	0	0	0	0	0	0	ENSG000000109016	DHRS7B	81	81	0	0	0	0	0	0	0
ENSG000000142544	CTU1	317	90	44	7	30	31	12	34	69	ENSG000000181192	DHTRD1	103	73	15	6	0	8	1	0	0
ENSG000000174177	CTU2	138	91	35	5	0	6	1	0	0	ENSG000000132153	DHX30	54	53	1	0	0	0	0	0	0
ENSG000000180891	CUEDC1	73	73	0	0	0	0	0	0	0	ENSG000000089876	DHX32	39	39	0	0	0	0	0	0	0
ENSG000000173681	CXorf23	42	42	0	0	0	0	0	0	0	ENSG000000184047	D1ABLO	52	52	0	0	0	0	0	0	0
ENSG000000166347	CYB5A	199	93	39	6	33	20	8	0	0	ENSG000000162946	DISC1	41	41	0	0	0	0	0	0	0
ENSG000000103018	CYB5B	36	36	0	0	0	0	0	0	0	ENSG000000150768	DLAT	259	94	36	6	33	20	8	11	51

EnsemblID	Gene	Sum	Hx	Hm	Ab	Ap	Sg	Ep	A	B
ENSG000000091140	DLD	283	97	36	6	33	33	11	10	57
ENSG000000119689	DLS1	247	95	36	6	32	30	8	0	40
ENSG000000132837	DMGDH	72	65	0	0	0	4	1	1	1
ENSG000000104936	DMPK	34	34	0	0	0	0	0	0	0
ENSG000000138346	DNA2	151	88	26	5	22	8	2	0	0
ENSG000000086061	DNAJA1	220	87	45	8	32	36	12	0	0
ENSG000000103423	DNAJA3	317	95	33	6	33	32	10	19	89
ENSG00000007923	DNAJC11	164	92	23	6	33	7	3	0	0
ENSG000000120675	DNAJC15	42	42	0	0	0	0	0	0	0
ENSG000000205981	DNAJC19	210	91	40	6	31	31	11	0	0
ENSG000000177692	DNAJC28	74	74	0	0	0	0	0	0	0
ENSG000000176410	DNAJC30	62	62	0	0	0	0	0	0	0
ENSG000000110011	DNAJC4	61	61	0	0	0	0	0	0	0
ENSG000000101152	DNAJC5	79	79	0	0	0	0	0	0	0
ENSG000000213221	DNLZ	196	84	35	6	31	33	7	0	0
ENSG000000087470	DNM1L	220	90	44	8	30	36	12	0	0
ENSG000000197959	DNM3	31	31	0	0	0	0	0	0	0
ENSG000000130816	DNMT1	68	68	0	0	0	0	0	0	0
ENSG000000119772	DNMT3A	63	63	0	0	0	0	0	0	0
ENSG000000107099	DOCK8	43	43	0	0	0	0	0	0	0
ENSG000000125170	DOK4	40	40	0	0	0	0	0	0	0
ENSG000000206052	DOK6	41	41	0	0	0	0	0	0	0
ENSG000000151640	DPYSL4	37	37	0	0	0	0	0	0	0
ENSG000000138101	DTNB	75	75	0	0	0	0	0	0	0
ENSG000000168393	DTYMK	305	96	44	7	33	34	11	0	80
ENSG000000167264	DUS2	168	94	37	7	25	5	0	0	0
ENSG000000167065	DUSP18	35	35	0	0	0	0	0	0	0
ENSG000000162999	DUSP19	111	75	1	6	24	5	0	0	0
ENSG000000189037	DUSP21	8	8	0	0	0	0	0	0	0
ENSG000000133878	DUSP26	44	44	0	0	0	0	0	0	0
ENSG000000128951	DUT	290	94	42	8	31	33	6	13	63
ENSG000000103356	EARS2	288	91	35	6	33	28	4	0	91
ENSG000000104823	ECHI	178	84	30	6	29	24	5	0	0
ENSG000000093144	ECHDC1	82	70	5	0	0	7	0	0	0
ENSG000000121310	ECHDC2	40	40	0	0	0	0	0	0	0
ENSG000000134463	ECHDC3	93	82	1	1	0	7	2	0	0
ENSG000000127884	ECHS1	237	92	30	6	10	23	8	14	54
ENSG000000167969	ECH2	97	70	0	0	9	10	8	0	0
ENSG000000198721	ECI2	103	87	0	1	0	14	1	0	0
ENSG000000130159	ECSIT	90	90	0	0	0	0	0	0	0
ENSG000000254772	EEF1G	198	94	34	6	24	30	10	0	0
ENSG000000115468	EFHD1	29	29	0	0	0	0	0	0	0
ENSG000000148730	EIF4EBP2	89	74	8	6	0	0	1	0	0
ENSG00000006744	ELAC2	219	92	45	8	32	30	12	0	0
ENSG000000126767	ELK1	60	60	0	0	0	0	0	0	0
ENSG000000104412	EMC2	194	94	27	8	31	26	8	0	0
ENSG000000131148	EMC8	179	91	19	6	29	26	8	0	0
ENSG000000167136	ENDOG	173	94	36	5	6	22	8	0	2
ENSG000000074800	ENO1	349	92	43	8	33	36	12	35	90
ENSG000000132199	ENOSF1	93	64	18	3	0	3	0	0	5
ENSG000000136628	EPRS	355	95	45	8	32	36	12	36	91
ENSG000000132591	ERAL1	247	92	7	6	33	26	8	0	75
ENSG000000182150	ERCC6L2	91	53	16	0	22	0	0	0	0
ENSG000000140009	ESR2	44	44	0	0	0	0	0	0	0
ENSG000000140374	ETFA	257	88	36	6	29	22	8	16	52
ENSG000000105379	ETFB	255	89	36	6	29	23	8	12	52
ENSG000000171503	ETFDH	203	90	36	6	29	24	8	0	10
ENSG000000105755	ETHE1	142	85	8	0	24	7	3	13	2
ENSG000000142459	EVI5L	36	36	0	0	0	0	0	0	0
ENSG000000081177	EXD2	97	73	0	5	3	7	9	0	0
ENSG000000157036	EXOG	39	39	0	0	0	0	0	0	0
ENSG000000117480	FAAH	114	71	29	1	1	10	2	0	0
ENSG000000149485	FADS1	35	35	0	0	0	0	0	0	0
ENSG000000180185	FAHDI	272	85	35	5	32	18	7	26	64
ENSG000000115042	FAHD2A	113	69	23	5	3	12	1	0	0
ENSG000000144199	FAHD2B	3	3	0	0	0	0	0	0	0
ENSG000000035141	FAM136A	135	89	0	0	27	19	0	0	0
ENSG000000114023	FAM162A	65	65	0	0	0	0	0	0	0
ENSG000000144369	FAM171B	43	43	0	0	0	0	0	0	0
ENSG000000103254	FAM173A	42	42	0	0	0	0	0	0	0
ENSG000000150756	FAM173B	96	83	0	0	0	10	3	0	0
ENSG000000222011	FAM185A	74	70	0	0	4	0	0	0	0
ENSG000000104059	FAM189A1	40	40	0	0	0	0	0	0	0
ENSG000000177150	FAM210A	79	79	0	0	0	0	0	0	0
ENSG000000124098	FAM210B	146	73	18	6	30	14	5	0	0
ENSG000000122378	FAM213A	62	58	4	0	0	0	0	0	0
ENSG000000139438	FAM222A	39	39	0	0	0	0	0	0	0
ENSG000000180488	FAM73A	43	43	0	0	0	0	0	0	0
ENSG000000148343	FAM73B	89	89	0	0	0	0	0	0	0
ENSG000000188343	FAM92A1	84	80	4	0	0	0	0	0	0
ENSG000000221829	FANCG	44	44	0	0	0	0	0	0	0
ENSG000000145982	FARS2	243	95	37	6	32	29	3	0	41
ENSG000000164896	FASTK	36	36	0	0	0	0	0	0	0
ENSG000000138399	FASTKDI	57	57	0	0	0	0	0	0	0

EnsemblID	Gene	Sum	Hs	Hm	Ab	Ap	Sg	Ep	A	B	EnsemblID	Gene	Sum	Hs	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000118246	<i>FASTKD2</i>	44	44	0	0	0	0	0	0	0	ENSG00000146729	<i>GBAS</i>	43	43	0	0	0	0	0	0	0
ENSG00000124279	<i>FASTKD3</i>	58	58	0	0	0	0	0	0	0	ENSG00000100116	<i>GCAT</i>	129	87	2	1	4	13	4	5	13
ENSG00000215251	<i>FASTKD5</i>	62	62	0	0	0	0	0	0	0	ENSG00000105607	<i>GCDD</i>	172	86	29	6	26	18	2	0	5
ENSG00000127452	<i>FBXL12</i>	53	53	0	0	0	0	0	0	0	ENSG00000140905	<i>GCSH</i>	288	91	37	6	33	25	10	16	70
ENSG00000112234	<i>FBXL4</i>	80	80	0	0	0	0	0	0	0	ENSG00000104381	<i>GDAP1</i>	64	44	0	0	18	2	0	0	0
ENSG00000100225	<i>FBXO7</i>	64	64	0	0	0	0	0	0	0	ENSG00000204183	<i>GDF5OS</i>	2	2	0	0	0	0	0	0	0
ENSG00000160752	<i>FDP5</i>	216	94	40	6	33	31	12	0	0	ENSG00000153982	<i>GDPD1</i>	115	76	21	0	2	11	5	0	0
ENSG00000137714	<i>FDX1</i>	92	84	0	0	0	6	2	0	0	ENSG00000127554	<i>GFER</i>	210	89	44	6	32	32	7	0	0
ENSG00000267673	<i>FDXIL</i>	215	88	41	6	33	35	11	0	1	ENSG00000168827	<i>GFM1</i>	306	96	37	5	33	32	8	4	91
ENSG00000161513	<i>FDXR</i>	220	90	40	6	32	35	8	0	9	ENSG00000164347	<i>GFM2</i>	131	85	22	5	0	11	7	0	1
ENSG00000066926	<i>FECH</i>	254	90	37	5	33	21	6	4	58	ENSG00000152964	<i>GGPS1</i>	145	86	34	6	1	15	3	0	0
ENSG00000168496	<i>FEN1</i>	265	95	45	8	33	35	12	37	0	ENSG00000165678	<i>GHTM</i>	111	92	16	1	0	2	0	0	0
ENSG0000011790	<i>FGFR1OP2</i>	92	92	0	0	0	0	0	0	0	ENSG00000152661	<i>GIA1</i>	45	45	0	0	0	0	0	0	0
ENSG00000091483	<i>FH</i>	254	93	37	6	31	19	4	15	49	ENSG00000198814	<i>GK</i>	250	92	36	8	33	23	12	5	41
ENSG00000189283	<i>FHIT</i>	175	53	40	5	31	28	11	0	7	ENSG00000196475	<i>GK2</i>	10	10	0	0	0	0	0	0	0
ENSG0000022267	<i>FHL1</i>	43	43	0	0	0	0	0	0	0	ENSG00000178445	<i>GLDC</i>	274	90	37	6	33	22	8	16	62
ENSG00000172500	<i>FIBP</i>	102	80	6	5	1	7	3	0	0	ENSG00000167699	<i>GLD4</i>	99	89	0	0	4	4	2	0	0
ENSG00000214253	<i>FISI</i>	173	85	36	4	30	16	2	0	0	ENSG00000171433	<i>GLD5</i>	71	52	7	3	1	1	1	0	6
ENSG00000141756	<i>FKBP10</i>	44	44	0	0	0	0	0	0	0	ENSG00000178445	<i>GLDC</i>	274	90	37	6	33	22	8	16	62
ENSG00000134285	<i>FKBP11</i>	39	39	0	0	0	0	0	0	0	ENSG0000023572	<i>GLRX2</i>	184	91	20	6	29	30	8	0	0
ENSG0000004478	<i>FKBP4</i>	91	91	0	0	0	0	0	0	0	ENSG00000182512	<i>GLRX5</i>	173	91	41	5	0	30	6	0	0
ENSG00000105701	<i>FKBP8</i>	86	86	0	0	0	0	0	0	0	ENSG00000115419	<i>GLS</i>	116	86	2	1	0	7	3	0	17
ENSG00000160688	<i>FLAD1</i>	107	78	1	0	28	0	0	0	0	ENSG00000148672	<i>GLS2</i>	30	30	0	0	0	0	0	0	0
ENSG00000162769	<i>FLVCR1</i>	95	84	0	6	0	5	0	0	0	ENSG00000182890	<i>GLUD1</i>	293	95	32	6	30	25	11	28	66
ENSG00000110074	<i>FOXRED1</i>	110	94	0	1	0	9	1	0	5	ENSG00000135821	<i>GLUL</i>	5	5	0	0	0	0	0	0	0
ENSG00000136877	<i>FPGS</i>	298	94	37	6	27	29	10	8	87	ENSG00000149124	<i>GLYAT</i>	197	95	31	1	31	21	8	0	10
ENSG00000188738	<i>FSP2</i>	25	25	0	0	0	0	0	0	0	ENSG00000168237	<i>GLYTK</i>	128	28	0	0	0	0	0	0	35
ENSG00000167996	<i>FTHL</i>	171	87	4	0	31	2	2	0	45	ENSG00000196743	<i>GM2A</i>	72	67	0	4	0	0	1	0	0
ENSG00000087086	<i>FTL</i>	28	28	0	0	0	0	0	0	0	ENSG00000197045	<i>GMEB</i>	116	83	25	8	0	0	0	0	0
ENSG00000181867	<i>FTMT</i>	21	21	0	0	0	0	0	0	0	ENSG00000204628	<i>GMB2L1</i>	225	94	45	8	33	34	11	0	0
ENSG00000122687	<i>FTSJ2</i>	181	87	15	2	4	26	4	23	20	ENSG00000174021	<i>GNG5</i>	40	40	0	0	0	0	0	0	0
ENSG00000069509	<i>FUNDC1</i>	144	88	16	5	8	21	3	0	3	ENSG00000113384	<i>GOLPH3</i>	135	96	39	0	0	0	0	0	0
ENSG00000165775	<i>FUNDC2</i>	31	31	0	0	0	0	0	0	0	ENSG00000125166	<i>GOT2</i>	206	91	34	6	31	33	11	0	0
ENSG00000165060	<i>FXN</i>	201	89	43	5	31	24	9	0	0	ENSG00000119927	<i>GP4M</i>	76	76	0	0	0	0	0	0	0
ENSG00000123689	<i>GOS2</i>	43	43	0	0	0	0	0	0	0	ENSG00000186281	<i>GP4T2</i>	33	33	0	0	0	0	0	0	0
ENSG00000179271	<i>GADD45GIP1</i>	76	76	0	0	0	0	0	0	0	ENSG00000186281	<i>GPDI</i>	253	94	43	1	24	33	9	0	49
ENSG00000156958	<i>GALK2</i>	167	94	30	6	30	7	0	0	0	ENSG00000115159	<i>GPD2</i>	252	95	43	5	30	33	8	1	37
ENSG00000111640	<i>GAPDH</i>	314	93	44	7	30	36	12	2	90	ENSG00000166123	<i>GP72</i>	211	95	36	8	31	25	12	0	4
ENSG00000106105	<i>GARS</i>	297	94	45	8	33	35	12	36	34	ENSG00000233276	<i>GPY1</i>	48	48	0	0	0	0	0	0	0
ENSG00000059691	<i>GATB</i>	295	87	35	5	30	24	1	28	85	ENSG00000167468	<i>GPY4</i>	227	85	44	3	30	23	9	0	33
ENSG00000257218	<i>GATC</i>	115	81	0	4	1	0	0	2	27	ENSG00000075240	<i>GRAMD4</i>	65	65	0	0	0	0	0	0	0
ENSG00000171766	<i>GATM</i>	71	59	0	0	1	8	1	0	2	ENSG00000137106	<i>GRHR</i>	163	80	15	5	5	8	4	11	35

EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000109519	GRPEL1	319	95	36	5	33	36	10	17	87
ENSG00000164284	GRPEL2	42	42	0	0	0	0	0	0	0
ENSG00000132463	GRSF1	41	41	0	0	0	0	0	0	0
ENSG00000131149	GSEI	57	57	0	0	0	0	0	0	0
ENSG00000104687	GSR	192	68	40	6	31	25	7	4	11
ENSG00000170899	GSTA4	39	39	0	0	0	0	0	0	0
ENSG00000197448	GSTK1	92	66	18	1	3	3	1	0	0
ENSG00000084207	GSTP1	54	54	0	0	0	0	0	0	0
ENSG00000100577	GSTZ1	145	84	12	6	26	15	2	0	0
ENSG00000105793	GTPBP10	90	90	0	0	0	0	0	0	0
ENSG00000130299	GTPBP3	269	80	30	6	32	27	4	0	90
ENSG00000178605	GTPBP6	237	83	0	1	33	18	2	21	79
ENSG00000163607	GTPBP8	225	61	25	6	33	30	8	0	62
ENSG00000151806	GUF1	293	85	36	7	33	33	8	0	91
ENSG00000143774	GUK1	310	92	44	8	32	34	11	0	89
ENSG00000138796	HADH	237	86	29	6	26	23	7	18	42
ENSG00000084754	HADHA	112	91	1	1	0	12	7	0	0
ENSG00000138029	HADHB	109	90	1	0	0	10	7	0	1
ENSG00000063854	HAGH	221	96	44	6	30	29	3	0	13
ENSG00000112855	HARS2	26	26	0	0	0	0	0	0	0
ENSG00000143575	HAX1	57	57	0	0	0	0	0	0	0
ENSG00000004961	HCCS	174	95	36	6	11	24	2	0	0
ENSG00000111906	HDDC2	201	92	35	8	33	22	11	0	0
ENSG00000119431	HDHD3	157	82	26	6	28	14	1	0	0
ENSG00000068097	HEATR6	115	68	7	6	26	6	2	0	0
ENSG00000013583	HEBP1	41	41	0	0	0	0	0	0	0
ENSG00000138411	HECW2	69	69	0	0	0	0	0	0	0
ENSG00000130589	HELZ2	49	49	0	0	0	0	0	0	0
ENSG00000114735	HEMK1	255	69	34	6	32	16	8	0	90
ENSG00000106049	HIBADH	176	87	32	6	23	20	3	0	5
ENSG00000198130	HIBCH	213	92	37	6	32	31	8	0	7
ENSG00000181061	HIGD1A	64	64	0	0	0	0	0	0	0
ENSG00000146066	HIGD2A	142	74	29	2	25	10	2	0	0
ENSG00000137133	HINT2	49	49	0	0	0	0	0	0	0
ENSG00000156515	HK1	44	44	0	0	0	0	0	0	0
ENSG00000159399	HK2	176	88	37	1	25	18	7	0	0
ENSG00000160883	HK3	26	26	0	0	0	0	0	0	0
ENSG00000156510	HKDC1	35	35	0	0	0	0	0	0	0
ENSG00000117305	HMGCL	187	84	28	6	28	18	3	0	20
ENSG00000134240	HMGCS2	32	32	0	0	0	0	0	0	0
ENSG00000165119	HNRNPK	76	76	0	0	0	0	0	0	0
ENSG00000241935	HOGA1	189	68	21	1	0	18	2	0	79
ENSG00000186603	HPDL	63	63	0	0	0	0	0	0	0
ENSG00000135116	HRK	22	22	0	0	0	0	0	0	0
ENSG00000132541	HRSP12	268	89	34	8	32	15	8	13	69
ENSG00000100209	HSCB	213	86	40	6	31	28	9	0	13
ENSG00000072506	HSD17B10	150	91	18	5	0	22	4	0	10
ENSG00000204228	HSD17B8	70	70	0	0	0	0	0	0	0
ENSG00000203859	HSD3B2	7	7	0	0	0	0	0	0	0
ENSG00000103160	HSDL1	69	69	0	0	0	0	0	0	0
ENSG00000119471	HSDL2	88	88	0	0	0	0	0	0	0
ENSG00000096384	HSP90AB1	40	40	0	0	0	0	0	0	0
ENSG00000166598	HSP90B1	129	89	5	7	24	4	0	0	0
ENSG00000204389	HSPA1A	24	24	0	0	0	0	0	0	0
ENSG00000044574	HSPA5	213	94	37	7	32	34	9	0	0
ENSG00000109971	HSPA8	219	94	40	8	32	33	12	0	0
ENSG00000113013	HSPA9	328	91	43	7	33	35	9	20	90
ENSG00000144381	HSPD1	313	96	37	8	33	36	11	2	90
ENSG00000115541	HSPE1	298	91	34	6	32	35	9	2	89
ENSG00000115317	HTRA2	231	79	3	0	32	21	1	12	83
ENSG00000149428	HYOU1	163	91	29	4	24	15	0	0	0
ENSG00000067704	IARS2	271	91	32	6	31	17	3	0	91
ENSG00000181873	IBA57	217	89	36	6	32	22	8	0	24
ENSG00000167862	ICT1	176	90	33	5	27	17	4	0	0
ENSG00000119912	IDE	218	91	45	6	33	30	8	0	5
ENSG00000138413	IDH1	79	79	0	0	0	0	0	0	0
ENSG00000182054	IDH2	198	80	35	6	33	27	6	0	11
ENSG00000166411	IDH3A	217	93	36	6	30	4	1	6	41
ENSG00000101365	IDH3B	77	77	0	0	0	0	0	0	0
ENSG00000067829	IDH3G	135	87	35	1	9	3	0	0	0
ENSG00000188483	IER5L	30	30	0	0	0	0	0	0	0
ENSG00000165949	IFI27	52	46	6	0	0	0	0	0	0
ENSG00000126709	IFI6	17	17	0	0	0	0	0	0	0
ENSG00000119917	IFIT3	23	23	0	0	0	0	0	0	0
ENSG00000105135	ILVBL	86	78	1	5	0	2	0	0	0
ENSG00000148950	IMMP1L	286	89	34	6	32	32	8	0	85
ENSG00000184903	IMMP2L	131	75	22	0	25	8	0	0	1
ENSG00000132305	IMMT	151	94	32	0	13	12	0	0	0
ENSG00000102794	IRG1	68	46	22	0	0	0	0	0	0
ENSG00000135070	ISCA1	260	93	37	6	32	31	8	7	46
ENSG00000165898	ISCA2	194	89	32	5	32	27	9	0	0
ENSG00000136003	ISCU	313	92	42	7	32	35	11	22	72
ENSG00000063241	ISOC2	61	61	0	0	0	0	0	0	0
ENSG00000128928	IVD	183	85	29	6	28	20	6	3	6

EnsemblID	Gene	Sum	Hs	Hm	Ab	Ap	Sg	Ep	A	B	EnsemblID	Gene	Sum	Hs	Hm	Ab	Ap	Sg	Ep	A	B
ENSG000000065427	KARS	319	95	45	8	33	36	11	6	85	ENSG000000120992	LYPLAI	205	91	39	6	31	30	8	0	0
ENSG000000151704	KCNJ1	44	44	0	0	0	0	0	0	0	ENSG000000102897	LYRM1	91	63	0	4	11	6	7	0	0
ENSG000000187486	KCNJ11	39	39	0	0	0	0	0	0	0	ENSG000000083099	LYRM2	74	66	8	0	0	0	0	0	0
ENSG000000156113	KCNMA1	61	61	0	0	0	0	0	0	0	ENSG000000214113	LYRM4	175	83	29	6	31	18	8	0	0
ENSG000000081791	KIAA0141	61	61	0	0	0	0	0	0	0	ENSG000000205707	LYRM5	91	76	6	6	0	2	1	0	0
ENSG000000100890	KIAA0391	131	83	2	0	29	11	6	0	0	ENSG000000186687	LYRM7	114	74	13	3	12	10	2	0	0
ENSG000000100364	KIAA0930	122	91	6	1	24	0	0	0	0	ENSG000000232859	LYRM9	84	60	0	0	24	0	0	0	0
ENSG000000122203	KIAA1191	44	44	0	0	0	0	0	0	0	ENSG000000140280	LYSMD2	84	80	0	4	0	0	0	0	0
ENSG000000163617	KIAA1407	66	66	0	0	0	0	0	0	0	ENSG000000159374	MIAP	60	60	0	0	0	0	0	0	0
ENSG000000117009	KIAA2013	85	85	0	0	0	0	0	0	0	ENSG000000183833	MAATSI	119	79	5	0	4	20	11	0	0
ENSG00000016685	KIF1B	41	41	0	0	0	0	0	0	0	ENSG000000133315	MACROD1	178	76	22	8	6	18	9	0	39
ENSG000000054523	KIF1B	41	41	0	0	0	0	0	0	0	ENSG000000156928	MAISU1	241	83	11	5	29	22	8	0	83
ENSG000000162755	KLHDC9	36	36	0	0	0	0	0	0	0	ENSG000000189221	MAOA	97	66	12	6	4	7	2	0	0
ENSG000000119771	KLHL29	44	44	0	0	0	0	0	0	0	ENSG000000069535	MAOB	31	31	0	0	0	0	0	0	0
ENSG000000167755	KLK6	21	21	0	0	0	0	0	0	0	ENSG000000100030	MAPK1	119	87	32	0	0	0	0	0	0
ENSG000000117009	KMO	166	83	33	6	25	15	4	0	0	ENSG000000185386	MAPK11	35	35	0	0	0	0	0	0	0
ENSG000000133703	KRAS	80	80	0	0	0	0	0	0	0	ENSG000000188130	MAPK12	40	40	0	0	0	0	0	0	0
ENSG000000186081	KRT5	24	24	0	0	0	0	0	0	0	ENSG000000102882	MAPK3	26	26	0	0	0	0	0	0	0
ENSG000000087299	L2HGDH	201	90	30	8	30	18	7	3	15	ENSG000000117791	MARCA2	162	80	17	6	30	14	2	0	13
ENSG000000135537	LACE1	212	87	35	6	33	34	8	0	9	ENSG000000186205	MARCI	21	21	0	0	0	0	0	0	0
ENSG000000103642	LACTB	164	80	17	5	3	10	1	4	44	ENSG000000198060	MARCH5	110	85	25	0	0	0	0	0	0
ENSG000000147592	LACTB2	134	87	25	0	3	15	4	0	0	ENSG000000247626	MARS2	210	90	36	6	33	34	11	0	0
ENSG000000135862	LAMC1	1	1	0	0	0	0	0	0	0	ENSG000000088888	MAVS	39	39	0	0	0	0	0	0	0
ENSG00000002549	LAP3	266	88	30	6	33	34	8	0	67	ENSG000000100294	MCAT	268	96	24	5	32	20	6	0	85
ENSG00000011376	LARS2	309	92	22	6	31	28	3	36	91	ENSG000000078070	MCCCI	174	85	28	6	24	14	5	0	12
ENSG000000166816	LDHD	188	73	30	5	28	15	6	14	17	ENSG000000131844	MCCC2	182	86	29	6	27	22	6	0	6
ENSG000000166816	LETM1	213	96	37	6	33	34	7	0	0	ENSG000000204511	MCCD1	22	22	0	0	0	0	0	0	0
ENSG000000165046	LETM2	41	41	0	0	0	0	0	0	0	ENSG000000124370	MCEE	112	78	7	6	5	9	7	0	0
ENSG000000050426	LETMD1	80	75	0	0	0	5	0	0	0	ENSG000000143384	MCCL1	48	48	0	0	0	0	0	0	0
ENSG000000121897	LIAS	266	97	37	3	33	24	8	8	56	ENSG000000156026	MCU	154	90	13	5	27	12	7	0	0
ENSG000000005156	LIG3	82	77	0	5	0	0	0	0	0	ENSG000000050393	MCURI	119	54	34	0	27	4	0	0	0
ENSG000000144182	LIP1T1	219	92	37	6	21	23	9	12	19	ENSG000000146701	MDH2	182	94	37	0	29	15	7	0	0
ENSG000000175536	LIP1T2	233	88	35	4	33	19	8	3	43	ENSG000000082212	ME2	42	42	0	0	0	0	0	0	0
ENSG000000074695	LMAN1	130	83	8	8	0	21	10	0	0	ENSG000000151376	ME3	40	40	0	0	0	0	0	0	0
ENSG000000196365	LONP1	239	93	33	6	30	31	3	4	39	ENSG000000116353	MECR	190	92	30	6	33	10	8	0	11
ENSG00000018095	LRRPRC	72	72	0	0	0	0	0	0	0	ENSG000000172878	METAP1D	107	71	4	6	21	5	0	0	0
ENSG000000198812	LRRC10	44	44	0	0	0	0	0	0	0	ENSG000000214756	METTL12	33	33	0	0	0	0	0	0	0
ENSG000000254402	LRRC24	45	45	0	0	0	0	0	0	0	ENSG000000169519	METTL15	255	82	7	6	32	28	9	0	91
ENSG000000108829	LRRC59	76	76	0	0	0	0	0	0	0	ENSG000000165792	METTL17	194	89	30	6	29	27	8	0	5
ENSG000000173988	LRRC63	53	53	0	0	0	0	0	0	0	ENSG000000139160	METTL20	64	64	0	0	0	0	0	0	0
ENSG000000154237	LRRK1	47	47	0	0	0	0	0	0	0	ENSG000000123600	METTL8	40	40	0	0	0	0	0	0	0
ENSG000000188906	LRRK2	85	85	0	0	0	0	0	0	0	ENSG000000197006	METTL9	113	89	1	6	3	7	7	0	0
ENSG000000160285	LSS	94	58	36	0	0	0	0	0	0											

EnsemblID	Gene	Sum	Hz	Hm	Ab	Sg	Ep	A	B	EnsemblID	Gene	Sum	Hz	Hm	Ab	Sg	Ep	A	B	
ENSG000000168958	MFF	70	70	0	0	0	0	0	0	ENSG000000115364	MRPL19	258	90	29	0	32	22	2	0	83
ENSG000000171109	MFN1	42	42	0	0	0	0	0	0	ENSG000000112651	MRPL2	272	89	36	1	31	20	4	0	91
ENSG000000116688	MFN2	92	92	0	0	0	0	0	0	ENSG000000242485	MRPL20	252	89	0	5	33	30	4	0	91
ENSG000000137463	MGARP	38	38	0	0	0	0	0	0	ENSG000000197345	MRPL21	207	91	21	6	31	27	8	0	23
ENSG000000125871	MGME1	90	90	0	0	0	0	0	0	ENSG00000082515	MRPL22	225	89	29	6	32	16	2	0	51
ENSG000000008394	MGST1	74	74	0	0	0	0	0	0	ENSG000000214026	MRPL23	172	83	32	4	25	15	6	0	7
ENSG000000143198	MGST3	157	75	27	1	33	19	2	0	ENSG000000143314	MRPL24	257	90	14	6	33	30	7	0	77
ENSG000000107745	MICU1	145	91	3	5	28	11	7	0	ENSG000000108826	MRPL27	282	83	35	6	33	31	3	0	91
ENSG000000165487	MICU2	42	42	0	0	0	0	0	0	ENSG000000086504	MRPL28	192	92	31	5	31	30	2	0	1
ENSG000000155970	MICU3	84	84	0	0	0	0	0	0	ENSG000000114686	MRPL3	302	93	38	7	33	32	8	0	91
ENSG000000100335	MIEF1	50	50	0	0	0	0	0	0	ENSG000000185414	MRPL30	98	87	1	0	1	0	1	0	8
ENSG000000177427	MIEF2	34	34	0	0	0	0	0	0	ENSG000000106591	MRPL32	118	86	23	3	2	4	0	0	0
ENSG000000173436	MINOS1	137	80	33	1	16	3	4	0	ENSG000000243147	MRPL33	164	61	32	5	28	21	2	0	15
ENSG0000000027001	MIPEP	193	94	34	6	33	19	7	0	ENSG000000130312	MRPL34	190	58	30	6	25	6	1	0	64
ENSG000000110917	MLEC	98	89	0	0	0	9	0	0	ENSG000000132313	MRPL35	88	77	1	0	10	0	0	0	0
ENSG000000108788	MLX	104	89	15	0	0	0	0	0	ENSG000000171421	MRPL36	183	58	23	2	26	12	1	0	61
ENSG000000175727	MLXIP	90	90	0	0	0	0	0	0	ENSG000000116221	MRPL37	87	87	0	0	0	0	0	0	0
ENSG000000103150	MLYCD	130	76	9	5	24	10	5	0	ENSG000000204316	MRPL38	107	94	12	1	0	0	0	0	0
ENSG000000151611	MMA4	170	77	7	6	4	12	5	18	ENSG000000154719	MRPL39	79	79	0	0	0	0	0	0	0
ENSG000000139428	MMAB	189	77	8	6	10	14	8	20	ENSG000000105364	MRPL4	293	92	36	5	33	29	7	0	91
ENSG000000132763	MMACHC	85	76	2	2	2	2	1	0	ENSG000000185608	MRPL40	85	85	0	0	0	0	0	0	0
ENSG000000168288	MMADHC	122	86	7	6	9	12	1	0	ENSG000000182154	MRPL41	138	81	11	1	24	20	1	0	0
ENSG000000168314	MOBP	24	24	0	0	0	0	0	0	ENSG000000198015	MRPL42	77	77	0	0	0	0	0	0	0
ENSG000000124615	MOCSI	261	91	21	5	33	17	5	29	ENSG0000000055950	MRPL43	178	89	27	5	30	24	3	0	0
ENSG000000115275	MOGS	174	89	32	6	30	15	2	0	ENSG000000135900	MRPL44	112	88	24	0	0	0	0	0	0
ENSG000000060762	MPC1	187	86	34	6	31	22	8	0	ENSG000000278845	MRPL45	138	91	16	1	29	1	0	0	0
ENSG000000143158	MPC2	194	88	35	6	32	26	7	0	ENSG000000259494	MRPL46	185	90	31	6	31	19	8	0	0
ENSG000000103152	MPG	146	69	0	0	22	5	3	6	ENSG000000136522	MRPL47	200	95	34	6	31	26	8	0	0
ENSG000000128309	MPST	223	69	34	6	33	28	10	11	ENSG000000175581	MRPL48	79	79	0	0	0	0	0	0	0
ENSG000000115204	MPV17	200	94	35	6	33	26	6	0	ENSG000000149792	MRPL49	151	87	31	6	4	20	3	0	0
ENSG000000254858	MPV17L2	87	76	11	0	0	0	0	0	ENSG000000136897	MRPL50	76	73	3	0	0	0	0	0	0
ENSG000000278619	MRM1	248	69	24	6	31	25	7	0	ENSG000000111639	MRPL51	78	78	0	0	0	0	0	0	0
ENSG000000169288	MRPL1	188	81	14	0	26	2	0	0	ENSG000000172590	MRPL52	71	71	0	0	0	0	0	0	0
ENSG000000159111	MRPL10	95	80	7	0	0	0	0	0	ENSG000000204822	MRPL53	89	71	3	1	13	0	1	0	0
ENSG000000174547	MRPL11	320	89	36	1	33	31	7	32	ENSG000000183617	MRPL54	140	87	11	0	30	9	3	0	0
ENSG000000262814	MRPL12	255	90	37	7	31	31	7	0	ENSG000000162910	MRPL55	79	79	0	0	0	0	0	0	0
ENSG000000172172	MRPL13	300	92	37	6	33	33	8	0	ENSG000000173141	MRPL57	69	69	0	0	0	0	0	0	0
ENSG000000180992	MRPL14	101	84	1	0	12	3	0	0	ENSG000000143436	MRPL9	202	81	4	6	28	11	1	0	71
ENSG000000137547	MRPL15	288	96	35	6	32	31	7	0	ENSG000000048544	MRPS10	108	93	15	0	0	0	0	0	0
ENSG000000166902	MRPL16	262	89	30	2	30	17	3	0	ENSG000000181991	MRPS11	233	91	20	0	29	6	0	0	87
ENSG000000158042	MRPL17	288	91	35	4	32	30	7	0	ENSG000000128626	MRPS12	266	86	36	1	29	23	0	0	91
ENSG000000112110	MRPL18	157	93	0	0	27	0	1	0	ENSG000000120333	MRPS14	246	81	34	1	27	15	0	0	88

EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B	EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000116898	<i>MRPS15</i>	246	78	30	2	30	21	3	0	82	ENSG00000198786	<i>MT-ND5</i>	236	86	12	1	30	9	1	22	75
ENSG00000182180	<i>MRPS16</i>	277	89	35	0	33	28	2	0	90	ENSG00000198695	<i>MT-ND6</i>	46	46	0	0	0	0	0	0	0
ENSG00000239789	<i>MRPS17</i>	217	86	24	4	29	16	6	0	52	ENSG00000137409	<i>MTCH1</i>	39	39	0	0	0	0	0	0	0
ENSG00000096080	<i>MRPS18A</i>	83	83	0	0	0	0	0	0	0	ENSG00000109919	<i>MTCH2</i>	91	91	0	0	0	0	0	0	0
ENSG00000204568	<i>MRPS18B</i>	81	81	0	0	0	0	0	0	0	ENSG00000214827	<i>MTCPI</i>	34	34	0	0	0	0	0	0	0
ENSG00000163319	<i>MRPS18C</i>	256	87	30	3	31	16	2	0	87	ENSG00000127989	<i>MTERF1</i>	38	38	0	0	0	0	0	0	0
ENSG00000122140	<i>MRPS2</i>	251	92	37	0	25	7	1	0	89	ENSG00000120832	<i>MTERF2</i>	44	44	0	0	0	0	0	0	0
ENSG00000266472	<i>MRPS21</i>	123	73	11	0	5	4	0	0	30	ENSG00000156469	<i>MTERF3</i>	127	87	0	0	30	9	1	0	0
ENSG00000175110	<i>MRPS22</i>	91	91	0	0	0	0	0	0	0	ENSG00000122085	<i>MTERF4</i>	61	61	0	0	0	0	0	0	0
ENSG00000181610	<i>MRPS23</i>	85	85	0	0	0	0	0	0	0	ENSG00000103707	<i>MTERMT</i>	202	85	35	6	19	19	6	0	32
ENSG00000062582	<i>MRPS24</i>	86	86	0	0	0	0	0	0	0	ENSG00000242114	<i>MTEP1</i>	128	80	27	1	6	11	3	0	0
ENSG00000131368	<i>MRPS25</i>	96	88	8	0	0	0	0	0	0	ENSG00000066855	<i>MTERL</i>	44	44	0	0	0	0	0	0	0
ENSG00000125901	<i>MRPS26</i>	74	74	0	0	0	0	0	0	0	ENSG00000117640	<i>MTERL1</i>	43	43	0	0	0	0	0	0	0
ENSG00000113048	<i>MRPS27</i>	80	80	0	0	0	0	0	0	0	ENSG00000146410	<i>MTER2</i>	66	66	0	0	0	0	0	0	0
ENSG00000147586	<i>MRPS28</i>	76	76	0	0	0	0	0	0	0	ENSG00000148824	<i>MTG1</i>	230	91	36	6	32	29	8	0	28
ENSG00000112996	<i>MRPS30</i>	85	85	0	0	0	0	0	0	0	ENSG00000101181	<i>MTG2</i>	292	91	32	6	33	32	8	0	90
ENSG00000102738	<i>MRPS31</i>	87	87	0	0	0	0	0	0	0	ENSG00000100714	<i>MTTHD1</i>	303	92	37	6	33	23	8	14	90
ENSG00000090263	<i>MRPS33</i>	97	87	1	0	9	0	0	0	0	ENSG00000120254	<i>MTTHD1L</i>	41	41	0	0	0	0	0	0	0
ENSG00000074071	<i>MRPS34</i>	105	78	5	5	8	5	4	0	0	ENSG00000065911	<i>MTTHF2</i>	67	65	0	2	0	0	0	0	0
ENSG00000061794	<i>MRPS35</i>	122	92	8	1	19	2	0	0	0	ENSG00000163738	<i>MTTHFD2L</i>	33	33	0	0	0	0	0	0	0
ENSG00000134056	<i>MRPS36</i>	58	55	2	1	0	0	0	0	0	ENSG00000136371	<i>MTHFS</i>	234	87	33	1	28	20	3	4	58
ENSG00000144029	<i>MRPS5</i>	177	90	35	1	31	17	3	0	0	ENSG00000085760	<i>MTIF2</i>	299	94	36	6	33	31	8	0	91
ENSG00000243927	<i>MRPS6</i>	146	84	28	0	11	8	2	0	13	ENSG00000122033	<i>MTIF3</i>	106	57	0	0	13	5	1	0	30
ENSG00000125445	<i>MRPS7</i>	246	90	29	1	30	4	1	0	91	ENSG00000135297	<i>MTI1</i>	266	76	31	6	33	29	3	0	88
ENSG00000135972	<i>MRPS9</i>	280	90	33	6	33	25	3	0	90	ENSG00000107951	<i>MTP4P</i>	73	73	0	0	0	0	0	0	0
ENSG00000148187	<i>MRRF</i>	258	81	24	5	32	23	3	0	90	ENSG00000120662	<i>MTRF1</i>	44	44	0	0	0	0	0	0	0
ENSG00000124532	<i>MRS2</i>	148	49	34	6	33	23	3	0	0	ENSG00000112031	<i>MTRF1L</i>	296	93	35	6	32	31	8	0	91
ENSG00000175806	<i>MSRA</i>	280	88	35	5	33	22	12	20	65	ENSG00000129422	<i>MTUS1</i>	44	44	0	0	0	0	0	0	0
ENSG00000148450	<i>MSRB2</i>	41	41	0	0	0	0	0	0	0	ENSG00000173171	<i>MTX1</i>	127	84	27	1	5	8	2	0	0
ENSG00000174099	<i>MSRB3</i>	270	93	34	6	33	25	10	15	54	ENSG00000128654	<i>MTX2</i>	74	74	0	0	0	0	0	0	0
ENSG00000125459	<i>MSTO1</i>	161	81	29	6	29	8	8	0	0	ENSG00000177034	<i>MTX3</i>	44	44	0	0	0	0	0	0	0
ENSG00000198899	<i>MT-ATP6</i>	184	77	16	1	26	6	0	0	58	ENSG00000090432	<i>MULL1</i>	74	74	0	0	0	0	0	0	0
ENSG00000228253	<i>MT-ATP8</i>	25	25	0	0	0	0	0	0	0	ENSG00000146085	<i>MUT</i>	180	80	7	6	4	14	6	19	44
ENSG00000198804	<i>MT-CO1</i>	215	92	17	2	26	18	3	12	45	ENSG00000132781	<i>MUTYH</i>	224	69	23	7	33	18	7	5	62
ENSG00000198712	<i>MT-CO2</i>	198	87	14	0	26	22	1	3	45	ENSG00000214114	<i>MYCBP</i>	113	80	5	0	0	16	12	0	0
ENSG00000198938	<i>MT-CO3</i>	182	86	15	1	26	14	1	0	39	ENSG00000133026	<i>MYH10</i>	130	82	40	8	0	0	0	0	0
ENSG00000198727	<i>MT-CYB</i>	205	86	17	1	26	19	1	10	45	ENSG00000100345	<i>MYH9</i>	40	40	0	0	0	0	0	0	0
ENSG00000198888	<i>MT-ND1</i>	231	85	14	1	30	11	1	24	65	ENSG00000106436	<i>MYL10</i>	30	30	0	0	0	0	0	0	0
ENSG00000198763	<i>MT-ND2</i>	154	71	10	1	19	8	0	9	36	ENSG00000278259	<i>MYO19</i>	44	44	0	0	0	0	0	0	0
ENSG00000198840	<i>MT-ND3</i>	177	77	8	1	29	9	0	0	53	ENSG00000036448	<i>MYOM2</i>	40	40	0	0	0	0	0	0	0
ENSG00000198886	<i>MT-ND4</i>	203	82	14	1	27	9	1	10	59	ENSG00000152620	<i>NADK2</i>	113	87	0	0	23	3	0	0	0
ENSG00000212907	<i>MT-ND4L</i>	79	61	2	0	6	5	0	0	5	ENSG00000161653	<i>NAGS</i>	55	50	0	5	0	0	0	0	0

EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000134265	NAPG	163	90	9	8	30	15	11	0	0
ENSG00000137513	NARS2	193	89	37	5	33	26	3	0	0
ENSG00000185818	NAT8L	54	50	4	0	0	0	0	0	0
ENSG00000166833	NAV2	43	43	0	0	0	0	0	0	0
ENSG00000188554	NBR1	81	53	0	6	20	2	0	0	0
ENSG00000125356	NDUFA1	66	51	7	0	6	2	0	0	0
ENSG00000130414	NDUFA10	85	85	0	0	0	0	0	0	0
ENSG00000174886	NDUFA11	64	56	8	0	0	0	0	0	0
ENSG00000184752	NDUFA12	187	93	32	6	31	18	3	0	4
ENSG00000186010	NDUFA13	178	89	30	6	32	13	8	0	0
ENSG00000131495	NDUFA2	169	83	31	6	33	14	2	0	0
ENSG00000170906	NDUFA3	47	41	0	0	3	3	0	0	0
ENSG00000189043	NDUFA4	91	71	18	0	1	1	0	0	0
ENSG00000185633	NDUFA4L2	26	26	0	0	0	0	0	0	0
ENSG00000128609	NDUFA5	177	91	31	6	29	18	2	0	0
ENSG00000184983	NDUFA6	169	89	31	4	28	10	7	0	0
ENSG00000267855	NDUFA7	89	87	2	0	0	0	0	0	0
ENSG00000119421	NDUFA8	155	94	32	0	29	0	0	0	0
ENSG00000139180	NDUFA9	224	94	33	6	33	18	8	4	28
ENSG00000004779	NDUFAB1	292	92	35	6	33	28	8	0	90
ENSG00000137806	NDUFAF1	174	93	29	5	31	14	2	0	0
ENSG00000164182	NDUFAF2	133	79	22	3	24	3	2	0	0
ENSG00000178057	NDUFAF3	170	87	26	6	29	15	7	0	0
ENSG00000123545	NDUFAF4	80	78	2	0	0	0	0	0	0
ENSG00000101247	NDUFAF5	228	92	31	6	33	16	8	0	42
ENSG00000156170	NDUFAF6	184	91	31	6	31	17	8	0	0
ENSG00000003509	NDUFAF7	216	93	32	6	32	16	8	0	29
ENSG00000183648	NDUFB1	52	52	0	0	0	0	0	0	0
ENSG00000140990	NDUFB10	94	84	3	0	6	0	1	0	0
ENSG00000147123	NDUFB11	76	73	3	0	0	0	0	0	0
ENSG00000090266	NDUFB2	81	77	0	1	3	0	0	0	0
ENSG00000119013	NDUFB3	104	78	26	0	0	0	0	0	0
ENSG00000065518	NDUFB4	71	71	0	0	0	0	0	0	0
ENSG00000136521	NDUFB5	83	83	0	0	0	0	0	0	0
ENSG00000165264	NDUFB6	57	57	0	0	0	0	0	0	0
ENSG00000099795	NDUFB7	159	86	25	6	30	10	2	0	0
ENSG00000166136	NDUFB8	120	90	21	5	4	0	0	0	0
ENSG00000147684	NDUFB9	178	91	32	6	29	13	7	0	0
ENSG00000109390	NDUFC1	35	35	0	0	0	0	0	0	0
ENSG00000151366	NDUFC2	63	63	0	0	0	0	0	0	0
ENSG0000023228	NDUFS1	235	96	32	1	33	12	6	0	55
ENSG00000023228	NDUFS2	253	93	33	1	25	9	2	24	66
ENSG000000213619	NDUFS3	233	94	33	2	24	9	1	15	55
ENSG00000164258	NDUFS4	183	93	33	6	30	17	2	0	2
ENSG00000168653	NDUFS5	100	76	7	0	12	5	0	0	0
ENSG00000145494	NDUFS6	172	93	33	6	24	15	1	0	0
ENSG00000115286	NDUFS7	275	95	33	6	32	14	6	21	68
ENSG00000110717	NDUFS8	264	91	33	5	32	17	1	24	61
ENSG00000167792	NDUFV1	258	96	34	6	32	18	10	4	58
ENSG00000178127	NDUFV2	259	97	34	6	32	19	10	3	58
ENSG00000160194	NDUFV3	56	56	0	0	0	0	0	0	0
ENSG00000204099	NEU4	38	38	0	0	0	0	0	0	0
ENSG00000244005	NFS1	332	97	44	8	33	36	11	15	88
ENSG00000169599	NFU1	253	94	36	5	33	33	10	0	42
ENSG00000165553	NGB	43	43	0	0	0	0	0	0	0
ENSG00000182768	NGRN	91	62	25	4	0	0	0	0	0
ENSG00000196290	NIF3L1	193	82	44	8	0	2	7	8	42
ENSG00000184117	NIPSNAP1	116	75	27	6	6	0	0	0	2
ENSG00000136783	NIPSNAP3A	54	54	0	0	0	0	0	0	0
ENSG00000165028	NIPSNAP3B	5	5	0	0	0	0	0	0	0
ENSG00000158793	NIT1	157	84	28	8	27	9	1	0	0
ENSG00000114021	NIT2	295	86	35	7	31	24	10	30	72
ENSG00000123213	NLN	35	35	0	0	0	0	0	0	0
ENSG00000160703	NLRX1	44	44	0	0	0	0	0	0	0
ENSG00000239672	NME1	27	27	0	0	0	0	0	0	0
ENSG00000243678	NME2	329	93	45	8	33	33	9	32	76
ENSG00000103024	NME3	37	37	0	0	0	0	0	0	0
ENSG00000103202	NME4	42	42	0	0	0	0	0	0	0
ENSG00000172113	NME6	89	76	9	4	0	0	0	0	0
ENSG00000163864	NMNAT3	37	37	0	0	0	0	0	0	0
ENSG00000112992	NNT	162	87	9	6	7	22	3	1	27
ENSG00000084092	NOA1	176	94	11	6	33	26	6	0	0
ENSG00000140939	NOL3	26	26	0	0	0	0	0	0	0
ENSG00000225921	NOL7	42	42	0	0	0	0	0	0	0
ENSG00000196943	NOP9	111	88	1	5	0	13	4	0	0
ENSG00000086991	NOX4	54	54	0	0	0	0	0	0	0
ENSG00000181019	NQO1	41	41	0	0	0	0	0	0	0
ENSG00000113580	NR3C1	43	43	0	0	0	0	0	0	0
ENSG00000178694	NSUN3	45	45	0	0	0	0	0	0	0
ENSG00000117481	NSUN4	94	88	6	0	0	0	0	0	0
ENSG00000168268	NT5DC2	42	42	0	0	0	0	0	0	0
ENSG0000011696	NT5DC3	126	80	2	6	21	14	3	0	0
ENSG00000205309	NT5M	38	38	0	0	0	0	0	0	0
ENSG00000065057	NTHL1	306	90	42	8	31	35	12	0	88

EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B	EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000151413	<i>NUBP1</i>	293	81	30	5	31	15	9	36	86	ENSG00000188582	<i>PAQR9</i>	41	41	0	0	0	0	0	0	0
ENSG00000070081	<i>NUCB2</i>	87	87	0	0	0	0	0	0	0	ENSG00000116288	<i>PARK7</i>	255	92	17	8	32	34	12	0	60
ENSG00000016621	<i>NUDT13</i>	43	43	0	0	0	0	0	0	0	ENSG00000175193	<i>P4RL</i>	177	92	35	6	26	15	3	0	0
ENSG000000213965	<i>NUDT19</i>	84	78	0	6	0	0	0	0	0	ENSG00000143799	<i>P4RP1</i>	151	91	14	5	21	13	7	0	0
ENSG00000164978	<i>NUDT2</i>	92	76	0	0	0	15	1	0	0	ENSG00000162396	<i>PARS2</i>	136	87	26	6	0	17	0	0	0
ENSG00000170917	<i>NUDT6</i>	126	74	0	6	28	13	5	0	0	ENSG00000173599	<i>PC</i>	181	87	32	3	8	17	4	2	28
ENSG00000167799	<i>NUDT8</i>	88	81	7	0	0	0	0	0	0	ENSG00000132570	<i>PCBD2</i>	63	63	0	0	0	0	0	0	0
ENSG00000170502	<i>NUDT9</i>	99	93	0	0	0	4	2	0	0	ENSG00000175198	<i>PCCA</i>	121	82	9	6	3	13	5	0	3
ENSG00000171773	<i>NXNLI</i>	37	37	0	0	0	0	0	0	0	ENSG00000114054	<i>PCCB</i>	169	85	12	6	5	14	6	0	41
ENSG00000089127	<i>OAS1</i>	40	39	0	1	0	0	0	0	0	ENSG00000100889	<i>PCK2</i>	83	81	2	0	0	0	0	0	0
ENSG00000065154	<i>OAT</i>	204	93	36	5	30	25	3	1	11	ENSG00000120265	<i>PCMT1</i>	272	94	27	6	33	27	9	25	51
ENSG00000109180	<i>OC1AD1</i>	75	75	0	0	0	0	0	0	0	ENSG00000174840	<i>PDE12</i>	124	86	0	2	5	24	7	0	0
ENSG00000145247	<i>OC1AD2</i>	37	37	0	0	0	0	0	0	0	ENSG00000186642	<i>PDE2A</i>	76	76	0	0	0	0	0	0	0
ENSG00000105953	<i>OGDH</i>	249	93	37	6	32	32	8	0	41	ENSG00000258429	<i>PDF</i>	38	38	0	0	0	0	0	0	0
ENSG00000197444	<i>OGDHL</i>	37	37	0	0	0	0	0	0	0	ENSG00000131828	<i>PDHA1</i>	244	97	43	5	32	27	8	1	31
ENSG00000114026	<i>OGG1</i>	220	89	40	5	30	24	6	16	10	ENSG00000163114	<i>PDHA2</i>	12	12	0	0	0	0	0	0	0
ENSG00000147162	<i>OGT</i>	164	91	23	6	31	13	0	0	0	ENSG00000168291	<i>PDHB</i>	244	95	42	6	33	25	8	1	34
ENSG00000162600	<i>OMAI</i>	198	69	36	6	30	13	2	0	42	ENSG00000110435	<i>PDHX</i>	68	68	0	0	0	0	0	0	0
ENSG00000198836	<i>OPAI</i>	96	96	0	0	0	0	0	0	0	ENSG00000167004	<i>PDIA3</i>	94	94	0	0	0	0	0	0	0
ENSG00000125741	<i>OPA3</i>	163	86	30	2	22	21	2	0	0	ENSG00000152256	<i>PDK1</i>	41	41	0	0	0	0	0	0	0
ENSG00000128694	<i>OSGEP1</i>	273	91	30	5	33	22	3	0	89	ENSG0000005882	<i>PDK2</i>	185	92	34	2	31	18	8	0	0
ENSG00000036473	<i>OTC</i>	255	59	33	0	33	14	6	32	78	ENSG00000067992	<i>PDK3</i>	42	42	0	0	0	0	0	0	0
ENSG00000165588	<i>OTX2</i>	83	82	0	0	0	1	0	0	0	ENSG0000004799	<i>PDK4</i>	33	33	0	0	0	0	0	0	0
ENSG00000155463	<i>OXAIL</i>	270	92	37	6	33	27	8	0	67	ENSG00000164951	<i>PDP1</i>	157	92	33	0	17	15	0	0	0
ENSG00000083720	<i>OXCT1</i>	169	86	24	6	0	16	7	0	30	ENSG00000172840	<i>PDP2</i>	41	41	0	0	0	0	0	0	0
ENSG00000198754	<i>OXCT2</i>	15	15	0	0	0	0	0	0	0	ENSG00000090857	<i>PDP3</i>	78	78	0	0	0	0	0	0	0
ENSG00000204237	<i>OXLD1</i>	87	64	20	1	0	2	0	0	0	ENSG00000148459	<i>PDSSI</i>	224	93	33	0	32	25	8	33	0
ENSG00000154814	<i>OXNAD1</i>	76	62	7	0	4	2	1	0	0	ENSG00000229833	<i>PET100</i>	76	64	12	0	0	0	0	0	0
ENSG00000164830	<i>OXRI</i>	210	93	39	5	31	33	9	0	0	ENSG00000247077	<i>PGAM5</i>	115	89	1	0	2	17	6	0	0
ENSG00000151093	<i>OXSM</i>	273	95	20	5	33	26	8	0	86	ENSG00000102144	<i>PGK1</i>	336	93	44	8	32	36	12	33	78
ENSG00000169860	<i>P2RY1</i>	44	44	0	0	0	0	0	0	0	ENSG00000164040	<i>PGRM2</i>	178	88	35	5	25	21	4	0	0
ENSG00000169313	<i>P2RY12</i>	43	43	0	0	0	0	0	0	0	ENSG00000087157	<i>PGSI</i>	186	94	43	6	5	32	6	0	0
ENSG00000122884	<i>P4HA1</i>	140	93	0	0	29	15	3	0	0	ENSG00000167085	<i>PGB</i>	212	95	35	6	33	35	8	0	0
ENSG00000185624	<i>P4HB</i>	213	95	44	5	28	32	9	0	0	ENSG00000215021	<i>PHB2</i>	224	89	37	6	33	27	6	8	18
ENSG00000174740	<i>PABPC5</i>	22	22	0	0	0	0	0	0	0	ENSG00000067177	<i>PHK1</i>	39	39	0	0	0	0	0	0	0
ENSG00000179364	<i>PACCS2</i>	91	89	0	0	0	0	2	0	0	ENSG00000165443	<i>PHYHIP1</i>	45	45	0	0	0	0	0	0	0
ENSG00000128050	<i>PALCS</i>	155	81	33	6	33	2	0	0	0	ENSG00000175309	<i>PHYKPL</i>	38	38	0	0	0	0	0	0	0
ENSG00000101349	<i>PAK7</i>	43	43	0	0	0	0	0	0	0	ENSG00000140451	<i>PIF1</i>	167	87	32	7	27	7	7	0	0
ENSG00000217930	<i>PAM16</i>	152	75	33	5	22	15	2	0	0	ENSG00000102309	<i>PIN4</i>	171	89	24	0	0	3	5	0	50
ENSG00000125779	<i>PANK2</i>	34	33	0	0	1	0	0	0	0	ENSG00000158828	<i>PINK1</i>	77	77	0	0	0	0	0	0	0
ENSG00000138801	<i>PAPSS1</i>	227	93	32	0	33	13	1	15	40											
ENSG00000198682	<i>PAPSS2</i>	41	41	0	0	0	0	0	0	0											

EnsemblID	Gene	Sum	Hx	Hm	Ab	Ap	Sg	Ep	A	B	Sum	Hx	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000241878	PISD	223	91	39	6	30	28	8	0	21	138	76	2	4	29	18	9	0	0
ENSG00000107959	PITRMI	203	84	24	8	33	30	11	0	13	78	78	0	0	0	0	0	0	0
ENSG00000067225	PKM	329	96	44	8	33	36	11	21	80	39	39	0	0	0	0	0	0	0
ENSG00000179598	PLD6	126	74	4	0	6	10	2	0	30	119	88	31	0	0	0	0	0	0
ENSG00000107020	PLGRKT	76	68	2	3	0	1	2	0	0	198	89	35	6	31	24	8	0	5
ENSG00000214456	PLIN5	23	23	0	0	0	0	0	0	0	36	36	0	0	0	0	0	0	0
ENSG00000102024	PLS3	172	95	37	6	22	9	3	0	0	296	90	36	8	33	33	10	1	85
ENSG00000141682	PMAIP1	33	33	0	0	0	0	0	0	0	43	43	0	0	0	0	0	0	0
ENSG00000165688	PMPCA	176	91	27	1	29	26	2	0	0	108	85	0	0	23	0	0	0	0
ENSG00000105819	PMPCB	298	95	38	6	33	35	10	3	78	152	90	10	8	31	9	4	0	0
ENSG00000127838	PNKD	53	53	0	0	0	0	0	0	0	222	92	45	7	33	33	12	0	0
ENSG0000039650	PNKP	207	89	39	8	33	31	7	0	0	104	76	1	0	9	13	5	0	0
ENSG0000006757	PNPLA4	53	53	0	0	0	0	0	0	0	101	76	0	0	0	17	8	0	0
ENSG00000135241	PNPLA8	131	77	0	0	33	18	3	0	0	62	62	0	0	0	0	0	0	0
ENSG00000108439	PNPO	172	95	29	5	8	5	1	0	29	83	83	0	0	0	0	0	0	0
ENSG00000138035	PNPT1	220	82	0	0	33	14	2	0	89	43	43	0	0	0	0	0	0	0
ENSG00000004142	POLDIP2	95	95	0	0	0	0	0	0	0	144	85	0	1	33	17	8	0	0
ENSG00000140521	POLG	133	96	37	0	0	0	0	0	0	131	90	4	0	21	10	3	0	3
ENSG00000256525	POLG2	69	69	0	0	0	0	0	0	0	88	88	0	0	0	0	0	0	0
ENSG00000099821	POLRMT	207	96	37	6	30	30	8	0	0	229	53	30	6	30	16	3	0	91
ENSG00000127948	POR	193	94	36	6	30	19	8	0	0	262	96	45	8	33	34	12	34	0
ENSG00000138777	PP42	38	38	0	0	0	0	0	0	0	198	93	11	6	3	5	1	15	64
ENSG00000166794	PP1B	179	92	33	6	32	6	10	0	0	210	92	44	8	32	29	5	0	0
ENSG00000108179	PP1F	210	91	41	6	33	31	8	0	0	224	91	15	8	32	34	12	32	0
ENSG00000163644	PPMIK	82	65	17	0	0	0	0	0	0	210	65	0	6	32	18	2	0	87
ENSG00000214517	PPME1	198	93	36	5	31	22	11	0	0	46	46	0	0	0	0	0	0	0
ENSG00000143224	PPOX	40	40	0	0	0	0	0	0	0	42	42	0	0	0	0	0	0	0
ENSG00000186298	PPP1CC	40	40	0	0	0	0	0	0	0	303	94	35	4	32	29	10	19	80
ENSG00000077157	PPP1R12B	77	77	0	0	0	0	0	0	0	24	24	0	0	0	0	0	0	0
ENSG00000167641	PPP1R14A	38	38	0	0	0	0	0	0	0	137	65	2	3	23	14	3	2	25
ENSG00000087074	PPP1R15A	49	49	0	0	0	0	0	0	0	98	47	0	0	20	8	1	0	22
ENSG00000182676	PPP1R27	45	45	0	0	0	0	0	0	0	135	93	10	6	2	18	6	0	0
ENSG00000196850	PPTC7	208	92	36	6	33	31	10	0	0	306	94	36	6	32	22	2	29	85
ENSG00000117450	PRDX1	47	47	0	0	0	0	0	0	0	291	94	28	8	30	31	12	2	86
ENSG00000167815	PRDX2	297	63	33	8	30	35	11	30	87	143	89	10	5	24	12	3	0	0
ENSG00000165672	PRDX3	64	64	0	0	0	0	0	0	0	38	38	0	0	0	0	0	0	0
ENSG00000123131	PRDX4	74	74	0	0	0	0	0	0	0	38	38	0	0	0	0	0	0	0
ENSG00000126432	PRDX5	162	69	34	6	32	18	3	0	0	114	83	5	8	0	12	6	0	0
ENSG00000117592	PRDX6	171	85	33	6	25	21	1	0	0	77	77	0	0	0	0	0	0	0
ENSG00000169230	PRELID1	117	93	24	0	0	0	0	0	0	8	8	0	0	0	0	0	0	0
ENSG00000186314	PRELID2	45	45	0	0	0	0	0	0	0	258	90	35	8	28	30	11	32	24
ENSG00000138078	PREPL	122	56	7	6	26	24	3	0	0	131	75	10	6	22	12	6	0	0

EnsemblID	Gene	Sum	H2	Hm	Ab	Ap	Sg	Ep	A	B	EnsemblID	Gene	Sum	H2	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000132155	<i>RAF1</i>	42	42	0	0	0	0	0	0	0	ENSG00000163684	<i>RPP14</i>	79	63	7	5	4	0	0	0	0
ENSG00000039560	<i>RAI14</i>	44	44	0	0	0	0	0	0	0	ENSG00000164587	<i>RPS14</i>	259	95	42	8	32	33	12	37	0
ENSG00000146282	<i>RARS2</i>	85	81	0	4	0	0	0	0	0	ENSG00000134419	<i>RPS154</i>	341	95	45	8	33	36	12	37	75
ENSG00000101546	<i>RBFA</i>	71	65	0	0	0	1	0	0	5	ENSG00000231500	<i>RPS18</i>	349	92	45	8	33	34	11	37	89
ENSG00000117906	<i>RCN2</i>	53	53	0	0	0	0	0	0	0	ENSG00000231500	<i>RPS3</i>	333	96	45	7	32	34	12	37	70
ENSG00000160439	<i>RDH13</i>	61	61	0	0	0	0	0	0	0	ENSG00000072133	<i>RPS6KA6</i>	40	40	0	0	0	0	0	0	0
ENSG00000240857	<i>RDH14</i>	50	50	0	0	0	0	0	0	0	ENSG00000108443	<i>RPS6KBI</i>	95	89	6	0	0	0	0	0	0
ENSG00000068615	<i>REEP1</i>	39	39	0	0	0	0	0	0	0	ENSG00000142937	<i>RPS8</i>	260	96	45	8	33	35	12	31	0
ENSG00000076043	<i>REXO2</i>	172	87	35	5	28	15	2	0	0	ENSG00000170889	<i>RPS9</i>	336	92	44	8	33	36	12	37	74
ENSG00000135002	<i>RFX</i>	278	92	36	6	27	31	11	0	75	ENSG00000168028	<i>RPSA</i>	262	94	45	7	32	35	12	37	0
ENSG00000067560	<i>RHOA</i>	140	95	45	0	0	0	0	0	0	ENSG00000166133	<i>RPU5D2</i>	300	95	43	8	33	35	11	0	75
ENSG00000126858	<i>RHOT1</i>	182	93	36	6	28	17	2	0	0	ENSG00000156990	<i>RPU5D3</i>	39	39	0	0	0	0	0	0	0
ENSG00000140983	<i>RHOT2</i>	36	36	0	0	0	0	0	0	0	ENSG00000165526	<i>RPU5D4</i>	148	81	10	6	29	15	7	0	0
ENSG00000164327	<i>RICTOR</i>	127	94	33	0	0	0	0	0	0	ENSG00000136444	<i>RSAD1</i>	216	63	8	6	32	14	3	1	89
ENSG00000166532	<i>RMKL8</i>	37	37	0	0	0	0	0	0	0	ENSG00000134321	<i>RSAD2</i>	64	54	4	1	1	3	1	0	0
ENSG00000117016	<i>RIMS3</i>	41	41	0	0	0	0	0	0	0	ENSG00000130363	<i>RSPH3</i>	127	82	3	0	9	21	12	0	0
ENSG00000129465	<i>RIPK3</i>	31	31	0	0	0	0	0	0	0	ENSG00000117616	<i>RSRP1</i>	35	35	0	0	0	0	0	0	0
ENSG00000176623	<i>RMND1</i>	90	90	0	0	0	0	0	0	0	ENSG00000114993	<i>RTKN</i>	69	69	0	0	0	0	0	0	0
ENSG00000137824	<i>RMND3</i>	57	57	0	0	0	0	0	0	0	ENSG00000130347	<i>RTN4IP1</i>	111	93	0	6	12	0	0	0	0
ENSG00000155906	<i>RMND1</i>	203	86	43	6	25	27	6	0	10	ENSG00000105784	<i>RUNDC3B</i>	43	43	0	0	0	0	0	0	0
ENSG00000171865	<i>RNASEH1</i>	214	86	33	0	13	16	9	0	57	ENSG00000183207	<i>RUVBL2</i>	219	89	41	8	33	36	12	0	0
ENSG00000135828	<i>RNASEL</i>	35	35	0	0	0	0	0	0	0	ENSG00000100347	<i>SAMM50</i>	213	94	38	5	31	2	0	0	43
ENSG00000137933	<i>RNF144B</i>	44	44	0	0	0	0	0	0	0	ENSG00000123453	<i>SARDH</i>	73	63	0	0	0	8	2	0	0
ENSG00000138942	<i>RNF185</i>	174	86	9	8	32	28	11	0	0	ENSG0000004139	<i>SARM1</i>	86	86	0	0	0	0	0	0	0
ENSG00000171861	<i>RNMTL1</i>	170	74	3	0	25	13	8	0	47	ENSG00000104835	<i>SARS2</i>	160	80	34	5	31	9	1	0	0
ENSG00000125995	<i>ROMO1</i>	130	79	34	4	11	2	0	0	0	ENSG00000141504	<i>SAT2</i>	161	67	20	5	21	9	2	17	20
ENSG00000153574	<i>RPL4</i>	268	94	43	6	33	26	2	32	32	ENSG00000143653	<i>SCCPDH</i>	194	84	34	1	31	25	7	4	8
ENSG00000147403	<i>RPL10</i>	226	93	45	8	33	35	12	0	0	ENSG00000133028	<i>SCO1</i>	249	94	36	6	33	32	8	6	34
ENSG00000198755	<i>RPL10A</i>	259	93	45	8	33	34	11	35	0	ENSG00000116171	<i>SCP2</i>	39	39	0	0	0	0	0	0	0
ENSG00000174748	<i>RPL15</i>	264	95	44	8	33	36	11	37	0	ENSG00000130489	<i>SDHA</i>	80	79	0	1	0	0	0	0	0
ENSG00000105640	<i>RPL18A</i>	227	96	45	7	33	33	12	1	0	ENSG00000073578	<i>SDHA</i>	284	97	36	5	33	34	8	28	43
ENSG00000100316	<i>RPL3</i>	262	95	45	8	33	34	11	36	0	ENSG00000205138	<i>SDHAF1</i>	105	49	20	5	23	8	0	0	0
ENSG00000156482	<i>RPL30</i>	251	96	44	8	32	33	11	27	0	ENSG00000167985	<i>SDHAF2</i>	179	90	37	6	21	15	5	0	5
ENSG00000109475	<i>RPL34</i>	236	92	43	6	33	35	11	16	0	ENSG00000196636	<i>SDHAF3</i>	154	81	36	6	9	17	5	0	0
ENSG00000182899	<i>RPL35A</i>	227	92	41	8	33	32	12	9	0	ENSG00000154079	<i>SDHAF4</i>	129	68	20	4	21	9	5	0	2
ENSG00000165502	<i>RPL36AL</i>	1	1	0	0	0	0	0	0	0	ENSG00000117118	<i>SDHB</i>	288	97	36	6	33	32	7	19	58
ENSG00000122406	<i>RPL5</i>	265	95	45	8	33	36	11	37	0	ENSG00000143252	<i>SDHC</i>	180	94	36	6	22	12	1	0	9
ENSG00000089009	<i>RPL6</i>	226	97	43	8	33	35	10	0	0	ENSG00000204370	<i>SDHD</i>	145	95	33	4	5	8	0	0	0
ENSG00000161016	<i>RPL8</i>	262	95	43	8	33	34	12	37	0	ENSG00000100445	<i>SDR39UI</i>	169	84	0	6	32	5	2	0	40
ENSG00000089157	<i>RPLP0</i>	261	95	45	8	33	35	12	33	0	ENSG00000187742	<i>SECISBP2</i>	40	40	0	0	0	0	0	0	0
ENSG00000177600	<i>RPLP2</i>	229	94	37	8	33	36	12	9	0	ENSG00000073169	<i>SELO</i>	185	75	34	1	32	21	3	1	18
ENSG00000118705	<i>RPN2</i>	167	91	24	6	28	17	1	0	0	ENSG00000108387	<i>SEPT4</i>	66	66	0	0	0	0	0	0	0

EnsemblID	Gene	Sum	Hx	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000122335	SEKAC1	125	89	0	3	29	4	0	0	0
ENSG00000185917	SETD4	127	82	34	3	8	0	0	0	0
ENSG00000155542	SETD9	67	62	0	5	0	0	0	0	0
ENSG00000164466	SFXN1	90	90	0	0	0	0	0	0	0
ENSG00000156398	SFXN2	66	66	0	0	0	0	0	0	0
ENSG00000107819	SFXN3	22	22	0	0	0	0	0	0	0
ENSG00000183605	SFXN4	35	35	0	0	0	0	0	0	0
ENSG00000144040	SFXN5	140	66	31	6	3	26	8	0	0
ENSG00000131370	SH3BP5	45	45	0	0	0	0	0	0	0
ENSG00000097033	SH3GLB1	86	86	0	0	0	0	0	0	0
ENSG00000160691	SHC1	82	82	0	0	0	0	0	0	0
ENSG00000182199	SHMT2	46	46	0	0	0	0	0	0	0
ENSG00000215475	SLAH3	33	33	0	0	0	0	0	0	0
ENSG00000142082	SIRT3	159	86	26	1	29	11	6	0	0
ENSG00000089163	SIRT4	62	62	0	0	0	0	0	0	0
ENSG00000124523	SIRT5	182	74	18	0	4	12	7	16	51
ENSG00000110911	SLC11A2	181	86	32	6	33	22	2	0	0
ENSG00000197208	SLC22A4	29	29	0	0	0	0	0	0	0
ENSG00000140090	SLC24A4	76	76	0	0	0	0	0	0	0
ENSG00000100075	SLC25A1	132	90	34	3	0	5	0	0	0
ENSG00000183048	SLC25A10	212	93	36	6	33	36	8	0	0
ENSG00000108528	SLC25A11	92	92	0	0	0	0	0	0	0
ENSG00000115840	SLC25A12	147	85	36	6	7	10	3	0	0
ENSG00000004864	SLC25A13	36	36	0	0	0	0	0	0	0
ENSG00000102078	SLC25A14	39	39	0	0	0	0	0	0	0
ENSG00000102743	SLC25A15	86	85	1	0	0	0	0	0	0
ENSG00000122912	SLC25A16	162	70	35	4	25	20	8	0	0
ENSG00000182902	SLC25A18	32	32	0	0	0	0	0	0	0
ENSG00000125454	SLC25A19	149	93	24	3	29	0	0	0	0
ENSG00000120329	SLC25A2	15	15	0	0	0	0	0	0	0
ENSG00000178537	SLC25A20	182	93	35	6	24	20	4	0	0
ENSG00000183032	SLC25A21	160	73	36	6	29	13	3	0	0
ENSG00000177542	SLC25A22	82	82	0	0	0	0	0	0	0
ENSG00000125648	SLC25A23	35	35	0	0	0	0	0	0	0
ENSG00000085491	SLC25A24	41	41	0	0	0	0	0	0	0
ENSG00000148339	SLC25A25	195	92	33	8	31	23	8	0	0
ENSG00000144741	SLC25A26	203	94	37	0	33	33	6	0	0
ENSG00000153291	SLC25A27	131	82	0	6	26	15	2	0	0
ENSG00000155287	SLC25A28	42	42	0	0	0	0	0	0	0
ENSG00000197119	SLC25A29	166	78	32	4	28	17	7	0	0
ENSG00000075415	SLC25A3	212	95	35	6	33	35	8	0	0
ENSG00000174032	SLC25A30	71	71	0	0	0	0	0	0	0
ENSG00000151475	SLC25A31	32	32	0	0	0	0	0	0	0
ENSG00000164933	SLC25A32	198	95	34	6	33	24	6	0	0
ENSG00000171612	SLC25A33	42	42	0	0	0	0	0	0	0
ENSG00000162461	SLC25A34	40	40	0	0	0	0	0	0	0
ENSG00000125434	SLC25A35	116	72	34	0	0	8	2	0	0
ENSG00000114120	SLC25A36	153	93	36	6	0	11	7	0	0
ENSG00000147454	SLC25A37	204	92	35	6	31	32	8	0	0
ENSG00000144659	SLC25A38	150	80	34	0	13	19	4	0	0
ENSG0000013306	SLC25A39	38	38	0	0	0	0	0	0	0
ENSG00000151729	SLC25A4	204	86	37	6	33	34	8	0	0
ENSG00000075303	SLC25A40	194	93	36	6	29	27	3	0	0
ENSG00000181240	SLC25A41	22	22	0	0	0	0	0	0	0
ENSG00000181035	SLC25A42	83	83	0	0	0	0	0	0	0
ENSG00000077713	SLC25A43	48	48	0	0	0	0	0	0	0
ENSG00000160785	SLC25A44	141	90	2	5	29	12	3	0	0
ENSG00000162241	SLC25A45	49	49	0	0	0	0	0	0	0
ENSG00000164209	SLC25A46	94	94	0	0	0	0	0	0	0
ENSG00000140107	SLC25A47	44	44	0	0	0	0	0	0	0
ENSG00000145832	SLC25A48	44	44	0	0	0	0	0	0	0
ENSG00000005022	SLC25A5	39	39	0	0	0	0	0	0	0
ENSG00000122696	SLC25A51	89	89	0	0	0	0	0	0	0
ENSG00000141437	SLC25A52	4	4	0	0	0	0	0	0	0
ENSG00000269743	SLC25A53	40	40	0	0	0	0	0	0	0
ENSG00000169100	SLC25A6	13	13	0	0	0	0	0	0	0
ENSG00000140284	SLC27A2	93	51	24	1	0	12	5	0	0
ENSG00000143554	SLC27A3	26	26	0	0	0	0	0	0	0
ENSG000000014824	SLC30A9	140	95	0	1	30	12	2	0	0
ENSG00000124786	SLC35B3	146	90	0	8	27	13	8	0	0
ENSG00000127526	SLC35E1	148	82	34	0	32	0	0	0	0
ENSG00000213699	SLC35F6	100	88	1	3	0	8	0	0	0
ENSG00000137700	SLC37A4	70	70	0	0	0	0	0	0	0
ENSG00000070214	SLC44A1	75	75	0	0	0	0	0	0	0
ENSG00000137968	SLC44A5	40	40	0	0	0	0	0	0	0
ENSG00000100678	SLC8A3	38	38	0	0	0	0	0	0	0
ENSG00000089060	SLC8B1	172	82	35	2	30	20	3	0	0
ENSG00000198689	SLC9A6	129	93	36	0	0	0	0	0	0
ENSG00000164038	SLC9B2	99	81	3	0	6	3	6	0	0
ENSG00000119705	SLIRP	58	58	0	0	0	0	0	0	0
ENSG00000184347	SLIT3	41	41	0	0	0	0	0	0	0
ENSG00000141391	SLMO1	36	36	0	0	0	0	0	0	0
ENSG00000101166	SLMO2	171	90	33	6	22	19	1	0	0
ENSG00000163206	SMCP	12	12	0	0	0	0	0	0	0

EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B	EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B
ENSG000000183172	<i>SMDT1</i>	73	73	0	0	0	0	0	0	0	ENSG000000213463	<i>SYNJ2BP</i>	40	40	0	0	0	0	0	0	0
ENSG000000163866	<i>SMIM12</i>	73	73	0	0	0	0	0	0	0	ENSG000000136463	<i>TACOI1</i>	255	89	27	0	33	13	3	0	90
ENSG000000250317	<i>SMIM20</i>	55	55	0	0	0	0	0	0	0	ENSG000000165632	<i>TAF3</i>	86	86	0	0	0	0	0	0	0
ENSG000000168273	<i>SMIM4</i>	56	56	0	0	0	0	0	0	0	ENSG000000186051	<i>TAL2</i>	40	40	0	0	0	0	0	0	0
ENSG000000198742	<i>SMURF1</i>	41	41	0	0	0	0	0	0	0	ENSG000000144559	<i>TAMM41</i>	195	94	37	6	31	24	3	0	0
ENSG000000099940	<i>SNAP29</i>	88	88	0	0	0	0	0	0	0	ENSG000000170921	<i>TANC2</i>	64	64	0	0	0	0	0	0	0
ENSG000000145335	<i>SNCA</i>	39	39	0	0	0	0	0	0	0	ENSG000000183597	<i>TANGO2</i>	160	85	28	5	28	8	1	0	5
ENSG000000197157	<i>SND1</i>	199	95	34	6	29	23	12	0	0	ENSG000000143374	<i>TARS2</i>	24	24	0	0	0	0	0	0	0
ENSG000000101298	<i>SNPH</i>	38	38	0	0	0	0	0	0	0	ENSG000000198650	<i>TAT</i>	151	85	9	6	28	17	6	0	0
ENSG000000142168	<i>SOD1</i>	190	93	31	6	26	18	1	1	14	ENSG000000102125	<i>TAZ</i>	42	42	0	0	0	0	0	0	0
ENSG000000112096	<i>SOD2</i>	313	97	45	8	33	35	11	19	65	ENSG000000146350	<i>TBC1D32</i>	83	72	2	0	1	6	2	0	0
ENSG000000163071	<i>SPATA18</i>	63	63	0	0	0	0	0	0	0	ENSG000000197226	<i>TBC1D9B</i>	41	41	0	0	0	0	0	0	0
ENSG000000166118	<i>SPATA19</i>	29	29	0	0	0	0	0	0	0	ENSG000000136270	<i>TBRG4</i>	65	65	0	0	0	0	0	0	0
ENSG00000006282	<i>SPATA20</i>	188	69	15	6	29	4	1	19	45	ENSG000000179152	<i>TCALM</i>	75	70	0	5	0	0	0	0	0
ENSG000000158792	<i>SPATA2L</i>	36	36	0	0	0	0	0	0	0	ENSG000000139437	<i>TCHP</i>	63	63	0	0	0	0	0	0	0
ENSG000000145375	<i>SPATA5</i>	102	63	26	6	0	7	0	0	0	ENSG000000110719	<i>TCIRG1</i>	29	29	0	0	0	0	0	0	0
ENSG000000118363	<i>SPCS2</i>	160	91	20	6	26	15	2	0	0	ENSG000000196116	<i>TDRD7</i>	60	60	0	0	0	0	0	0	0
ENSG000000133104	<i>SPG20</i>	80	80	0	0	0	0	0	0	0	ENSG000000182134	<i>TDRKH</i>	70	70	0	0	0	0	0	0	0
ENSG000000197912	<i>SPG7</i>	78	78	0	0	0	0	0	0	0	ENSG000000172171	<i>TEFM</i>	51	51	0	0	0	0	0	0	0
ENSG000000153820	<i>SPHKAP</i>	44	44	0	0	0	0	0	0	0	ENSG000000164362	<i>TERT</i>	151	72	21	4	24	23	7	0	0
ENSG000000169682	<i>SPNS1</i>	137	79	0	0	27	6	0	4	21	ENSG000000108064	<i>TFAM</i>	71	71	0	0	0	0	0	0	0
ENSG000000116096	<i>SPR</i>	118	89	11	5	0	13	0	0	0	ENSG000000029639	<i>TFB1M</i>	102	88	7	6	0	1	0	0	0
ENSG000000176422	<i>SPRYD4</i>	42	42	0	0	0	0	0	0	0	ENSG000000162851	<i>TFB2M</i>	56	56	0	0	0	0	0	0	0
ENSG000000162032	<i>SPSB3</i>	76	76	0	0	0	0	0	0	0	ENSG000000072274	<i>TFRC</i>	44	44	0	0	0	0	0	0	0
ENSG000000100596	<i>SPTL2</i>	206	94	44	6	33	18	11	0	0	ENSG000000041988	<i>THAP3</i>	27	27	0	0	0	0	0	0	0
ENSG000000137767	<i>SQRDL</i>	169	90	21	5	0	10	6	15	22	ENSG000000159445	<i>THEM4</i>	135	47	34	5	11	20	2	1	15
ENSG000000068784	<i>SRBD1</i>	143	84	0	1	4	2	5	0	47	ENSG000000196407	<i>THEM5</i>	26	26	0	0	0	0	0	0	0
ENSG000000197122	<i>SRC</i>	43	43	0	0	0	0	0	0	0	ENSG000000113272	<i>THGIL</i>	183	87	37	6	29	20	4	0	0
ENSG000000106028	<i>SSBP1</i>	214	74	18	1	24	11	2	0	84	ENSG000000185875	<i>THNSL1</i>	50	50	0	0	0	0	0	0	0
ENSG000000147465	<i>STAR</i>	43	43	0	0	0	0	0	0	0	ENSG000000134809	<i>TIMM10</i>	189	90	34	6	31	22	6	0	0
ENSG000000133121	<i>STAR13</i>	66	66	0	0	0	0	0	0	0	ENSG000000132286	<i>TIMM10B</i>	61	61	0	0	0	0	0	0	0
ENSG000000084090	<i>STAR17</i>	138	85	0	3	27	21	2	0	0	ENSG000000099800	<i>TIMM13</i>	183	86	35	1	30	28	3	0	0
ENSG000000168610	<i>STAR17</i>	42	42	0	0	0	0	0	0	0	ENSG000000134375	<i>TIMM17A</i>	42	42	0	0	0	0	0	0	0
ENSG000000165283	<i>STOML2</i>	281	94	22	3	31	31	8	30	62	ENSG000000126768	<i>TIMM17B</i>	208	92	37	6	32	35	6	0	0
ENSG000000136874	<i>STX17</i>	68	68	0	0	0	0	0	0	0	ENSG000000075336	<i>TIMM21</i>	146	88	22	1	25	9	1	0	0
ENSG000000136143	<i>SUCL42</i>	297	91	36	6	33	31	10	31	59	ENSG000000177370	<i>TIMM22</i>	200	90	40	6	33	25	6	0	0
ENSG000000163541	<i>SUCLG1</i>	303	95	36	6	33	33	10	31	59	ENSG0000000265354	<i>TIMM23</i>	163	90	34	2	24	12	1	0	0
ENSG000000172340	<i>SUCLG2</i>	95	85	0	5	0	5	0	0	0	ENSG000000204152	<i>TIMM23B</i>	66	41	0	2	21	2	0	0	0
ENSG000000175600	<i>SUGCT</i>	177	80	27	6	8	15	2	12	27	ENSG000000104980	<i>TIMM44</i>	197	94	37	3	32	26	3	0	2
ENSG000000139531	<i>SUXO</i>	149	86	20	5	5	17	3	0	13	ENSG000000105197	<i>TIMM50</i>	180	87	39	0	32	20	2	0	0
ENSG000000156502	<i>SUP3L1</i>	187	81	36	6	29	28	7	0	0	ENSG000000126953	<i>TIMM84</i>	156	86	31	0	30	8	1	0	0
ENSG000000148290	<i>SURF1</i>	203	95	34	5	33	19	8	0	9	ENSG000000150779	<i>TIMM8B</i>	41	41	0	0	0	0	0	0	0

EnsemblID	Gene	Sum	Hx	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000100575	TIMM9	183	81	36	5	30	26	5	0	0
ENSG00000113845	TIMMDC1	80	80	0	0	0	0	0	0	0
ENSG00000166548	TK2	80	80	0	0	0	0	0	0	0
ENSG00000101337	TM9SF4	194	87	33	6	32	26	10	0	0
ENSG00000139644	TM6IM6	163	90	5	6	30	24	8	0	0
ENSG00000091947	TM6IM101	55	55	0	0	0	0	0	0	0
ENSG00000178307	TM6IM11	88	88	0	0	0	0	0	0	0
ENSG00000171202	TM6IM1264	58	58	0	0	0	0	0	0	0
ENSG00000171204	TM6IM126B	26	26	0	0	0	0	0	0	0
ENSG00000244187	TM6IM141	48	48	0	0	0	0	0	0	0
ENSG00000161558	TM6IM143	53	53	0	0	0	0	0	0	0
ENSG00000111843	TM6IM14C	157	78	28	5	30	11	3	0	2
ENSG00000130748	TM6IM160	42	42	0	0	0	0	0	0	0
ENSG00000144120	TM6IM177	66	66	0	0	0	0	0	0	0
ENSG00000184857	TM6IM186	71	71	0	0	0	0	0	0	0
ENSG00000168569	TM6IM223	75	66	4	0	0	5	0	0	0
ENSG00000165152	TM6IM246	61	61	0	0	0	0	0	0	0
ENSG00000205544	TM6IM256	135	80	6	6	22	10	2	0	9
ENSG00000137038	TM6IM261	38	38	0	0	0	0	0	0	0
ENSG00000164983	TM6IM65	126	88	1	0	11	19	7	0	0
ENSG00000159596	TM6IM69	80	55	14	0	11	0	0	0	0
ENSG00000175606	TM6IM70	94	79	2	1	12	0	0	0	0
ENSG00000185973	TM6IM77	115	74	30	1	0	8	2	0	0
ENSG00000133687	TM6IM7C1	52	51	0	0	0	0	1	0	0
ENSG00000139324	TM6IM7C3	75	75	0	0	0	0	0	0	0
ENSG00000213593	TM6IM7C2	130	90	4	5	21	9	1	0	0
ENSG00000173726	TM6IM7C2	120	92	28	0	0	0	0	0	0
ENSG00000196860	TM6IM7C2L	27	27	0	0	0	0	0	0	0
ENSG00000100216	TM6IM7C2	117	89	6	0	20	2	0	0	0
ENSG00000025772	TM6IM7C3	45	39	0	0	1	5	0	0	0
ENSG00000130204	TM6IM7C4	195	91	39	3	32	28	2	0	0
ENSG00000158882	TM6IM7C4L	34	34	0	0	0	0	0	0	0
ENSG00000175768	TM6IM7C5	43	42	1	0	0	0	0	0	0
ENSG00000214736	TM6IM7C6	43	43	0	0	0	0	0	0	0
ENSG00000196683	TM6IM7C7	85	58	18	0	3	5	1	0	0
ENSG00000154174	TM6IM7C7A	126	90	36	0	0	0	0	0	0
ENSG00000184428	TM6IM7C7	18	18	0	0	0	0	0	0	0
ENSG00000177302	TM6IM7C7A	347	95	43	7	30	34	12	35	91
ENSG00000120471	TP53AIP1	8	8	0	0	0	0	0	0	0
ENSG00000111669	TPH1	326	95	44	8	33	29	11	19	87
ENSG00000166340	TPP1	71	51	0	6	1	11	2	0	0
ENSG00000170638	TRABD	164	88	1	6	31	11	3	21	3
ENSG00000175104	TRAF6	65	60	0	5	0	0	0	0	0
ENSG00000126602	TRAP1	193	94	6	6	4	29	7	0	47
ENSG00000170855	TRAP1	156	86	33	6	19	12	0	0	0
ENSG00000155890	TRIM42	36	36	0	0	0	0	0	0	0
ENSG00000043514	TRIT1	303	97	36	8	31	33	9	0	89
ENSG00000174173	TRMT10C	75	75	0	0	0	0	0	0	0
ENSG00000066651	TRMT11	223	96	42	8	30	35	12	0	0
ENSG00000188917	TRMT2B	34	34	0	0	0	0	0	0	0
ENSG00000126814	TRMT5	263	95	45	7	33	36	12	35	0
ENSG00000171103	TRMT61B	56	46	10	0	0	0	0	0	0
ENSG00000100416	TRMTU	275	89	26	6	32	24	7	0	91
ENSG00000072756	TRNT1	309	95	42	8	33	35	9	0	87
ENSG00000100991	TRPC4AP	84	63	0	4	6	10	1	0	0
ENSG00000167112	TRUB2	82	82	0	0	0	0	0	0	0
ENSG00000123297	TSFM	280	90	20	6	33	32	8	0	91
ENSG00000100300	TSP0	165	81	22	1	28	7	3	6	17
ENSG00000128311	TST	35	35	0	0	0	0	0	0	0
ENSG00000011295	TTC19	88	88	0	0	0	0	0	0	0
ENSG00000183891	TTC32	41	41	0	0	0	0	0	0	0
ENSG00000167552	TUBA1A	71	71	0	0	0	0	0	0	0
ENSG00000196230	TUBB	38	38	0	0	0	0	0	0	0
ENSG00000258947	TUBB3	31	31	0	0	0	0	0	0	0
ENSG00000178952	TUFM	300	93	37	7	33	31	8	0	91
ENSG00000114383	TUSC2	63	63	0	0	0	0	0	0	0
ENSG00000104723	TUSC3	172	91	37	5	27	11	1	0	0
ENSG00000100348	TXN2	254	95	13	5	29	11	3	21	77
ENSG00000117862	TXNDC12	76	75	0	1	0	0	0	0	0
ENSG00000184470	TXNRD2	65	65	0	0	0	0	0	0	0
ENSG00000025708	TYMP	111	62	0	0	0	4	1	0	44
ENSG00000176890	TYMS	267	97	44	1	32	36	8	15	34
ENSG00000130985	UBA1	223	94	44	8	30	35	12	0	0
ENSG00000186591	UBE2H	205	96	43	6	30	23	7	0	0
ENSG00000177889	UBE2N	214	97	36	5	33	33	10	0	0
ENSG00000151148	UBE3B	171	96	6	0	32	29	8	0	0
ENSG00000120942	UBIAD1	116	70	0	5	3	9	1	0	28
ENSG00000186150	UBL4B	25	25	0	0	0	0	0	0	0
ENSG00000109424	UCP1	20	20	0	0	0	0	0	0	0
ENSG00000175567	UCP2	98	62	0	0	29	7	0	0	0
ENSG00000175564	UCP3	35	35	0	0	0	0	0	0	0
ENSG00000060600	UHRF1BP1	41	41	0	0	0	0	0	0	0
ENSG0000011647	UHRF1BP1L	115	88	0	6	21	0	0	0	0
ENSG00000076248	UNG	246	86	44	8	30	31	10	0	37

EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B	EnsemblID	Gene	Sum	Hz	Hm	Ab	Ap	Sg	Ep	A	B
ENSG00000101019	<i>UQCC1</i>	171	87	33	6	30	13	2	0	0	ENSG00000028116	<i>VRK2</i>	43	43	0	0	0	0	0	0	0
ENSG00000137288	<i>UQCC2</i>	72	72	0	0	0	0	0	0	0	ENSG00000102763	<i>VWA8</i>	113	80	14	1	0	11	7	0	0
ENSG00000204922	<i>UQCC3</i>	36	36	0	0	0	0	0	0	0	ENSG00000116874	<i>WARS2</i>	285	89	35	6	32	29	3	0	91
ENSG00000184076	<i>UQCR10</i>	140	72	30	4	29	4	1	0	0	ENSG00000112290	<i>WASF1</i>	43	43	0	0	0	0	0	0	0
ENSG00000127540	<i>UQCR11</i>	56	55	0	1	0	0	0	0	0	ENSG00000274523	<i>WBSCL16</i>	98	90	8	0	0	0	0	0	0
ENSG00000156467	<i>UQCRB</i>	183	92	37	6	32	8	8	0	0	ENSG00000148225	<i>WDR31</i>	75	62	0	0	1	7	5	0	0
ENSG0000010256	<i>UQCRC1</i>	40	40	0	0	0	0	0	0	0	ENSG00000152763	<i>WDR78</i>	118	79	5	0	8	17	9	0	0
ENSG00000140740	<i>UQCRC2</i>	96	84	11	0	0	1	0	0	0	ENSG00000167716	<i>WDR81</i>	125	89	6	6	24	0	0	0	0
ENSG00000169021	<i>UQCRFS1</i>	237	94	35	6	33	25	8	0	36	ENSG00000179314	<i>WSCD1</i>	46	41	5	0	0	0	0	0	0
ENSG00000173660	<i>UQCRH</i>	166	88	29	2	29	16	2	0	0	ENSG00000196236	<i>WSCP2</i>	69	69	0	0	0	0	0	0	0
ENSG00000164405	<i>UQCRQ</i>	119	89	22	4	1	3	0	0	0	ENSG00000132530	<i>XAF1</i>	35	35	0	0	0	0	0	0	0
ENSG00000173915	<i>USMG5</i>	54	54	0	0	0	0	0	0	0	ENSG00000126215	<i>XRC3</i>	100	59	12	5	17	6	1	0	91
ENSG00000135093	<i>USP30</i>	116	83	16	0	17	0	0	0	0	ENSG00000139131	<i>YARS2</i>	286	90	35	6	33	28	3	0	78
ENSG00000115464	<i>USP34</i>	91	71	2	6	0	9	3	0	0	ENSG00000167645	<i>YBEY</i>	175	54	3	0	18	16	6	0	0
ENSG00000106346	<i>USP42</i>	44	44	0	0	0	0	0	0	0	ENSG00000182362	<i>YFI1B</i>	211	89	44	8	33	27	10	0	0
ENSG00000166348	<i>USP54</i>	71	71	0	0	0	0	0	0	0	ENSG00000250067	<i>YIEFN3</i>	40	40	0	0	0	0	0	0	0
ENSG00000147679	<i>UTP23</i>	179	91	29	5	27	25	2	0	0	ENSG00000136758	<i>YME1L1</i>	197	89	34	5	32	29	8	0	0
ENSG00000139190	<i>VAMP1</i>	39	39	0	0	0	0	0	0	0	ENSG00000196449	<i>YRDC</i>	261	94	1	7	32	30	12	0	85
ENSG00000204394	<i>VARS</i>	349	92	45	8	33	36	12	32	91	ENSG00000108953	<i>YWHAE</i>	212	81	45	8	33	33	12	0	0
ENSG00000137411	<i>VARS2</i>	40	40	0	0	0	0	0	0	0	ENSG00000134308	<i>YWHAQ</i>	34	34	0	0	0	0	0	0	0
ENSG00000108828	<i>VATI</i>	43	43	0	0	0	0	0	0	0	ENSG00000164924	<i>YWHAZ</i>	76	76	0	0	0	0	0	0	0
ENSG00000213585	<i>VDAC1</i>	44	44	0	0	0	0	0	0	0	ENSG00000180011	<i>ZADH2</i>	100	58	5	0	25	5	7	0	0
ENSG00000165637	<i>VDAC2</i>	176	94	36	3	29	11	3	0	0	ENSG00000221886	<i>ZBED8</i>	19	19	0	0	0	0	0	0	0
ENSG00000078668	<i>VDAC3</i>	36	36	0	0	0	0	0	0	0	ENSG00000099904	<i>ZDHHHC8</i>	42	42	0	0	0	0	0	0	0
ENSG00000026025	<i>VIM</i>	45	45	0	0	0	0	0	0	0	ENSG00000074755	<i>ZZEF1</i>	55	55	0	0	0	0	0	0	0
ENSG00000197969	<i>VPS134</i>	92	92	0	0	0	0	0	0	0											

Appendix II: Mitochondrial genes matching mitochondrial carrier phylogenetic patterns in nematodes and platyhelminthes

Table 1. IMPI 2017 positive training set genes matching the phylogenetic pattern of the SLC25A21 and SLC25A43 transporters in the studied platyhelminthes – present in the free-living *Schmidtea mediterranea* and absent in the blood flukes, liver flukes and tapeworms.

Ensembl ID	Gene	Function
ENSG00000109576	<i>AADAT</i>	Transamination of 2-aminoadipate to 2-oxoadipate and kynurenine to kynurenic acid
ENSG00000183044	<i>ABAT</i>	Conversion of GABA to succinate semialdehyde
ENSG00000167315	<i>ACAA2</i>	Acetyl-CoA acyltransferase – last step of mitochondria fatty acid beta-oxidation spiral
ENSG00000167107	<i>ACSF2</i>	Acyl-CoA synthetase, preferring medium-chain substrates
ENSG00000154930	<i>ACSS1</i>	Acetyl-CoA synthetase
ENSG00000147576	<i>ADHFE1</i>	Oxidation of γ -hydroxybutyrate to succinate semialdehyde
ENSG00000113492	<i>AGXT2</i>	Multifunctional aminotransferase
ENSG00000164904	<i>ALDH7A1</i>	Production of betaine aldehyde from betaine and 2-aminoadipate from 2-aminoadipate 6-semialdehyde
ENSG00000174606	<i>ANGEL2</i>	Unknown
ENSG00000125375	<i>ATP5S</i>	Subunit of ATP synthase
ENSG00000189227	<i>C15orf61</i>	Unknown
ENSG00000145439	<i>CBR4</i>	Quinone reductase and 3-keto-acyl-acyl carrier protein reductase in fatty acid biosynthesis
ENSG00000121289	<i>CEP89</i>	Complex IV regulation
ENSG00000181192	<i>DHTKD1</i>	Production of glutaryl-CoA from 2-oxoadipate
ENSG00000132837	<i>DMGDH</i>	Demethylation of dimethylglycine to sarcosine
ENSG00000132199	<i>ENOSF1</i>	Catabolism of <i>L</i> -fucose to 2-keto-3-deoxy- <i>L</i> -fucose
ENSG00000130299	<i>GTPBP3</i>	Mitochondrial tRNA modification
ENSG00000151806	<i>GUF1</i>	Mitoribosome fidelity factor – catalyses movement of ribosome back one codon
ENSG00000181061	<i>HIGD1A</i>	Complex IV regulation
ENSG00000189221	<i>MAOA</i>	Outer mitochondrial membrane monoamine metabolism
ENSG00000050393	<i>MCUR1</i>	Regulation of calcium uniporter
ENSG00000103150	<i>MLYCD</i>	Catalyses conversion of malonyl-CoA to acetyl-CoA
ENSG00000103152	<i>MPG</i>	DNA repair
ENSG00000254858	<i>MPV17L2</i>	Regulation of mitoribosome assembly and stability
ENSG00000108826	<i>MRPL27</i>	Subunit of the mitoribosome
ENSG00000135297	<i>MTO1</i>	Mitoribosome tRNA modification

Ensembl ID	Gene	Function
ENSG00000161653	<i>NAGS</i>	Production of <i>N</i> -acetylglutamate – a cofactor for <i>CPS1</i> in the urea cycle
ENSG00000162600	<i>OMA1</i>	Mitochondrial quality control
ENSG00000036473	<i>OTC</i>	Conversion of carbamoyl phosphate and <i>L</i> -ornithine to <i>L</i> -citrulline and P _i , in the urea cycle
ENSG00000128050	<i>PAICS</i>	Catalyses two steps in purine biosynthesis
ENSG00000179598	<i>PLD6</i>	Possibly cardiolipin metabolism
ENSG00000138035	<i>PNPT1</i>	RNA processing
ENSG00000126432	<i>PRDX5</i>	Reactive oxygen species metabolism
ENSG00000148334	<i>PTGES2</i>	Prostaglandin E2 synthase
ENSG00000169972	<i>PUSL1</i>	Unknown
ENSG00000118508	<i>RAB32</i>	Regulation of protein kinase A signalling and mitochondrial fission
ENSG00000136444	<i>RSAD1</i>	Unknown
ENSG00000123453	<i>SARDH</i>	Production of glycine from sarcosine
ENSG00000122912	<i>SLC25A16</i>	Groups with <i>SLC25A42</i> – the CoA transporter
ENSG00000183032	<i>SLC25A21</i>	Exchange of mitochondrial 2-oxoglutarate for cytoplasmic 2-oxoadipate
ENSG00000077713	<i>SLC25A43</i>	Unknown
ENSG00000070214	<i>SLC44A1</i>	Choline transporter
ENSG00000106028	<i>SSBP1</i>	Single-stranded DNA binding protein involved in mitochondrial DNA replication
ENSG00000175600	<i>SUGCT</i>	Conversion of glutarate to glutaryl-CoA
ENSG00000179152	<i>TCAIM</i>	Regulation of T-cell activity
ENSG00000164362	<i>TERT</i>	Maintenance of telomeres
ENSG00000185973	<i>TMLHE</i>	Conversion of trimethyllysine to hydroxytrimethyllysine
ENSG00000128311	<i>TST</i>	Cyanide detoxification
ENSG00000102763	<i>VWA8</i>	Unknown

Table 2. IMPI 2017 positive training set genes matching the phylogenetic pattern of the SLC25A38 transporter in the studied platyhelminthes – present in the free-living *Schmidtea mediterranea*, liver flukes and tapeworms, and absent in the blood flukes.

Ensembl ID	Gene	Function
ENSG00000115657	<i>ABCB6</i>	Transport of a porphyrin across the membrane in haem synthesis
ENSG00000125246	<i>CLYBL</i>	Production of malate from glyoxylate and acetyl-CoA; and β -methylmalate from glyoxylate and propionyl-CoA
ENSG00000157184	<i>CPT2</i>	Production of <i>N</i> -acyl-CoA from imported <i>N</i> -acylcarnitine
ENSG00000127884	<i>ECHS1</i>	Hydration of 2-trans-enoyl-CoAs to <i>L</i> -3-hydroxyacyl-CoAs in mitochondrial fatty acid β -oxidation
ENSG00000168237	<i>GLYCK</i>	Glycerate kinase
ENSG00000138796	<i>HADH</i>	Oxidation of straight chain 2-hydroxyacyl-CoAs in mitochondrial fatty acid β -oxidation
ENSG00000005156	<i>LIG3</i>	DNA ligase involved in repair
ENSG00000266472	<i>MRPS21</i>	Subunit of the mitoribosome
ENSG00000151413	<i>NUBPL</i>	Complex I assembly factor
ENSG00000142082	<i>SIRT3</i>	Lysine deacetylase
ENSG00000085491	<i>SLC25A24</i>	ADP/ATP carrier
ENSG00000144659	<i>SLC25A38</i>	Probable glycine carrier
ENSG00000175606	<i>TMEM70</i>	ATP synthase assembly factor

Table 3. IMPI 2017 positive training set genes matching the phylogenetic pattern of the SLC25A1 and SLC25A29 transporters in the studied nematodes – missing from the *Dorylaimia*.

Ensembl ID	Gene	Function
ENSG00000135776	<i>ABCB10</i>	Possibly related to haem transport
ENSG00000143149	<i>ALDH9A1</i>	Conversion of γ -aminobutyraldehyde to γ -butyrobetaine
ENSG00000256053	<i>APOPT1</i>	Regulation of apoptosis
ENSG00000139637	<i>C12orf10</i>	Unknown
ENSG00000016391	<i>CHDH</i>	Production of betaine aldehyde from choline
ENSG00000006695	<i>COX10</i>	Produces haem O from protohaem IX, for usage in Complex IV
ENSG00000149485	<i>FADS1</i>	Fatty acid desaturase
ENSG00000104687	<i>GSR</i>	Reduction of oxidised glutathione disulphide
ENSG00000119431	<i>HDHD3</i>	Unknown
ENSG00000067829	<i>IDH3G</i>	Regulatory subunit of isocitrate dehydrogenase
ENSG00000184903	<i>IMMP2L</i>	Removal of targeting sequence
ENSG00000103642	<i>LACTB</i>	Possibly involved in lipid metabolism or as a filament protein
ENSG00000166816	<i>LDHD</i>	Lactate dehydrogenase
ENSG00000005156	<i>LIG3</i>	DNA ligase involved in repair
ENSG00000108829	<i>LRRC59</i>	Regulation of mitochondrial dynamics?
ENSG00000156928	<i>MALSU1</i>	Regulation of mitoribosome activity
ENSG00000103150	<i>MLYCD</i>	Production of acetyl-CoA from malonyl-CoA
ENSG00000175806	<i>MSRA</i>	Repair of damaged methionine residues in proteins
ENSG00000134265	<i>NAPG</i>	Trafficking between organelles
ENSG00000158828	<i>PINK1</i>	Kinase involved in stress response
ENSG00000107959	<i>PITRM1</i>	Degradation of cleaved targeting sequences
ENSG00000101546	<i>RBFA</i>	Maturation of mitoribosome
ENSG00000126858	<i>RHOT1</i>	Mitochondrial trafficking
ENSG00000196636	<i>SDHAF3</i>	Complex II assembly factor
ENSG00000144040	<i>SFXN5</i>	Unknown
ENSG00000110911	<i>SLC11A2</i>	Metal ion transporter (particularly iron)
ENSG00000100075	<i>SLC25A1</i>	Citrate transporter
ENSG00000197119	<i>SLC25A29</i>	Basic amino acid transporter
ENSG00000084090	<i>STARD7</i>	Phosphatidylcholine trafficking

Table 4. IMPI 2017 positive training set genes matching the phylogenetic pattern of the SLC25A14 and SLC25A30 transporters in the studied platyhelminthes – present in the free-living *Schmidtea mediterranea* and liver flukes; missing in the blood flukes and tapeworms.

Ensembl ID	Gene	Function
ENSG00000008311	AASS	First two steps of lysine degradation, producing α -amino adipic semialdehyde
ENSG00000115361	ACADL	Long chain acyl-CoA dehydrogenase
ENSG00000122971	ACADS	Short chain acyl-CoA dehydrogenase
ENSG00000196177	ACADSB	Short branched chain acyl-CoA dehydrogenase
ENSG00000072778	ACADVL	Very long chain acyl-CoA dehydrogenase
ENSG00000111058	ACSS3	Acyl-CoA synthetase – required for fatty acid synthesis
ENSG00000119711	ALDH6A1	Production of acetyl-CoA and propionyl-CoA from malonate and methylmalonate respectively
ENSG00000148090	AUH	Conversion of 3-methylglutaconyl-CoA to 3-hydroxy-3-methylglutaryl-CoA
ENSG00000248098	BCKDHA	Branched chain amino acid dehydrogenase subunit
ENSG00000083123	BCKDHB	Branched chain amino acid dehydrogenase subunit
ENSG00000103507	BCKDK	Control of branched chain amino acid dehydrogenase
ENSG00000213563	C8orf82	Unknown
ENSG00000173992	CCS	Chaperone for copper insertion into superoxide dismutase
ENSG00000110090	CPT1A	Formation of acylcarnitine
ENSG00000137992	DBT	Branched chain amino acid dehydrogenase subunit
ENSG00000104325	DECR1	Accessory protein of mitochondrial fatty acid β -oxidation, dealing with unsaturated fatty acids
ENSG00000104823	ECH1	Enoyl-CoA hydratase
ENSG00000198721	ECI2	Accessory protein of mitochondrial fatty acid β -oxidation, dealing with unsaturated fatty acids
ENSG00000140374	EFTA	Electron transfer flavoprotein subunit
ENSG00000105379	ETFB	Electron transfer flavoprotein subunit
ENSG00000171503	ETFDH	Electron transfer flavoprotein dehydrogenase
ENSG00000105607	GCDH	Decarboxylation of glutaryl-CoA to crotonyl-CoA
ENSG00000105607	GSTK1	Glutathione transferase
ENSG00000100577	GSTZ1	Conversion of maleylacetoacetate to fumarylacetoacetate
ENSG00000084754	HADHA	Last steps of mitochondrial fatty acid β -oxidation
ENSG00000138029	HADHB	Last steps of mitochondrial fatty acid β -oxidation
ENSG00000106049	HIBADH	Oxidation of 3-hydroxyisobutyrate to methylmalonate semialdehyde
ENSG00000198130	HIBCH	Hydrolysis of 3-hydroxyisobutyryl-CoA and 3-hydroxypropanoyl-CoA
ENSG00000117305	HMGCL	Final step of leucine catabolism and start of ketone body synthesis
ENSG00000119471	HSDL2	Unknown
ENSG00000128928	IVD	Conversion of isovaleryl-CoA to 3-methylcrotonyl-CoA
ENSG00000078070	MCCC1	Methylcrotonyl-CoA carboxylase

Ensembl ID	Gene	Function
ENSG00000131844	<i>MCCC2</i>	Methylcrotonyl-CoA carboxylase
ENSG00000151611	<i>MMAA</i>	Vitamin B12 metabolism
ENSG00000139428	<i>MMAB</i>	Vitamin B12 metabolism
ENSG00000146085	<i>MUT</i>	Vitamin B12 dependent production of succinyl-CoA from methylmalonyl-CoA
ENSG00000213965	<i>NUDT19</i>	Renal CoA diphosphohydrolase
ENSG00000170917	<i>NUDT6</i>	Unknown – possibly control of cell proliferation
ENSG00000083720	<i>OXCT1</i>	Ketone body catabolism
ENSG00000163644	<i>PPM1K</i>	Control of branched chain amino acid dehydrogenase
ENSG00000187024	<i>PTRH1</i>	Unknown
ENSG00000182199	<i>SHMT2</i>	Interconversion of glycine and serine
ENSG00000089163	<i>SIRT4</i>	Removal of biotinyl and lipoyl groups
ENSG00000174032	<i>SLC25A30</i>	Unknown

Table 5. IMPI 2017 positive training set genes matching the phylogenetic pattern of the SLC25A43 transporter in the studied nematodes – only in the *Dorylaimia*.

Ensembl ID	Gene	Function
ENSG00000178074	<i>C2orf69</i>	Unknown
ENSG00000162129	<i>CLPB</i>	Peptidase
ENSG00000140465	<i>CYP1A1</i>	Wide-ranging monooxygenase
ENSG00000172500	<i>FIBP</i>	Regulation of mitogenesis
ENSG00000115317	<i>HTRA2</i>	Protease linked to apoptosis
ENSG00000002549	<i>LAP3</i>	Aminopeptidase
ENSG00000108788	<i>MLX</i>	Sensing cellular energy status and transcription regulator
ENSG00000103152	<i>MPG</i>	DNA glycosylase involved in repair
ENSG00000125459	<i>MSTO1</i>	Regulation of mitochondrial morphology
ENSG00000107951	<i>MTPAP</i>	Addition of polyA tail to mitochondrial mRNA
ENSG00000109180	<i>OCIAD1</i>	Unknown
ENSG00000125741	<i>OPA3</i>	Unknown
ENSG00000106246	<i>PTCD1</i>	Mitochondrial tRNA regulation
ENSG00000049883	<i>PTCD2</i>	Mitochondrial tRNA regulation
ENSG00000077713	<i>SLC25A43</i>	Unknown
ENSG00000113272	<i>THG1L</i>	Mitochondrial tRNA modification

Table 6. IMPI 2017 positive training set genes matching the phylogenetic pattern of the SLC25A44 transporter in the studied platyhelminthes – missing in the tapeworms.

Ensembl ID	Gene	Function
ENSG00000059573	<i>ALDH18A1</i>	Conversion of glutamate to glutamate 5-semialdehyde
ENSG00000143149	<i>ALDH9A1</i>	Conversion of γ -aminobutyraldehyde to γ -butyrobetaine
ENSG00000145020	<i>AMT</i>	Part of glycine cleavage system
ENSG00000256053	<i>APOPT1</i>	Regulation of apoptosis
ENSG00000174928	<i>C3orf33</i>	Unknown
ENSG00000016391	<i>CHDH</i>	Production of betaine aldehyde from choline
ENSG00000167113	<i>COQ4</i>	Organisation of part of the coenzyme Q biosynthesis pathway
ENSG00000102967	<i>DHODH</i>	Production of orotate from dihydroorotate with quinone as an electron acceptor
ENSG00000180185	<i>FAHD1</i>	Oxaloacetate decarboxylase
ENSG00000091483	<i>FH</i>	Fumarase
ENSG00000171766	<i>GATM</i>	Synthesis of guanidinoacetate from arginine and glycine
ENSG00000178445	<i>GLDC</i>	Part of glycine cleavage system
ENSG00000101365	<i>IDH3B</i>	Subunit of isocitrate dehydrogenase
ENSG00000087299	<i>L2HGDH</i>	Oxidisation of <i>L</i> -2-hydroxyglutarate to α -ketoglutarate
ENSG00000135537	<i>LACE1</i>	Mitochondrial morphology and protein metabolism
ENSG00000124615	<i>MOCS1</i>	First step of molybdenum cofactor synthesis
ENSG00000174099	<i>MSRB3</i>	Reduction of methionine sulfoxide to methionine
ENSG00000103707	<i>MTFMT</i>	Addition of formyl group to initiator methionyl-tRNA
ENSG00000214114	<i>MYCBP</i>	Control of <i>MYC</i> transcription factor
ENSG00000123213	<i>NLN</i>	Oligopeptidase hydrolysis
ENSG00000173599	<i>PC</i>	Pyruvate carboxylase
ENSG00000186642	<i>PDE2A</i>	cAMP and cGMP phosphodiesterase
ENSG00000164306	<i>PRIMPOL</i>	DNA primase and polymerase
ENSG00000111737	<i>RAB35</i>	Regulation of trafficking
ENSG00000171861	<i>RNMTL1</i>	Mitochondrial rRNA modification
ENSG00000160785	<i>SLC25A44</i>	Unknown
ENSG00000213699	<i>SLC35F6</i>	Mitochondrial membrane potential maintenance
ENSG00000163071	<i>SPATA18</i>	Mitochondrial quality control
ENSG00000139531	<i>SUOX</i>	Oxidation of sulphite to sulphate

Appendix III: Mitochondrial proteins in viruses

Table 1. IMPI 2017 genes with at least one predicted viral orthologue, with information about the gene family and the families of viruses with predicted orthologues.

Ensembl gene ID	Gene	HUGO gene family	Viral orthologue count	Viral families with orthologues
ENSG00000128951	<i>DUT</i>		745	Adenoviridae, Alloherpesviridae, Asfarviridae, Baculoviridae, Caulimoviridae, Herpesviridae, Hytrosaviridae, Iridoviridae, Marseilleviridae, Mimiviridae, Myoviridae, Nimaviridae, Nudiviridae, Pandoraviridae, Phycodnaviridae, Podoviridae, Poxviridae, Retroviridae, Siphoviridae, Unclassified
ENSG00000176890	<i>TYMS</i>		531	Alloherpesviridae, Herpesviridae, Hytrosaviridae, Iridoviridae, Marseilleviridae, Mimiviridae, Myoviridae, Nimaviridae, Nudiviridae, Phycodnaviridae, Podoviridae, Poxviridae, Siphoviridae, Unclassified
ENSG00000106028	<i>SSBP1</i>		339	Myoviridae, Podoviridae, Siphoviridae, Unclassified
ENSG00000125656	<i>CLPP</i>	AAA ATPases	305	Herpesviridae, Myoviridae, Phycodnaviridae, Podoviridae, Siphoviridae, Unclassified
ENSG00000099821	<i>POLRMT</i>		297	Myoviridae, Podoviridae, Siphoviridae, Unclassified
ENSG00000171865	<i>RNASEH1</i>		262	Caulimoviridae, Iridoviridae, Marseilleviridae, Mimiviridae, Myoviridae, Pandoraviridae, Phycodnaviridae, Retroviridae, Siphoviridae, Unclassified
ENSG00000076248	<i>UNG</i>	DNA glycosylases	203	Alloherpesviridae, Herpesviridae, Marseilleviridae, Mimiviridae, Pandoraviridae, Phycodnaviridae, Podoviridae, Poxviridae, Unclassified
ENSG00000142168	<i>SOD1</i>		192	Baculoviridae, Mimiviridae, Phycodnaviridae, Poxviridae, Unclassified
ENSG00000107815	<i>C10orf2</i>	DNA helicases	155	Myoviridae, Podoviridae, Siphoviridae, Unclassified

Ensembl gene ID	Gene	HUGO gene family	Viral orthologue count	Viral families with orthologues
ENSG00000127554	<i>GFER</i>		132	Ascoviridae, Asfarviridae, Iridoviridae, Marseilleviridae, Mimiviridae, Pandoraviridae, Phycodnaviridae, Unclassified
ENSG00000023572	<i>GLRX2</i>	Glutaredoxin domain containing	103	Mimiviridae, Myoviridae, Phycodnaviridae, Siphoviridae, Unclassified
ENSG00000168496	<i>FEN1</i>		101	Ascoviridae, Iridoviridae, Marseilleviridae, Mimiviridae, Phycodnaviridae, Unclassified
ENSG00000168393	<i>DTYMK</i>		97	Alloherpesviridae, Asfarviridae, Baculoviridae, Iridoviridae, Myoviridae, Phycodnaviridae, Podoviridae, Poxviridae, Siphoviridae, Unclassified
ENSG00000140451	<i>PIF1</i>	DNA helicases	72	Ascoviridae, Baculoviridae, Iridoviridae, Myoviridae, Pandoraviridae, Phycodnaviridae, Unclassified
ENSG00000100348	<i>TXN2</i>		72	Marseilleviridae, Myoviridae, Phycodnaviridae, Podoviridae, Siphoviridae, Unclassified
ENSG00000051180	<i>RAD51</i>	Fanconi anemia complementation groups	67	Myoviridae, Siphoviridae, Unclassified
ENSG00000072756	<i>TRNT1</i>		61	Bicaudaviridae, Myoviridae, Podoviridae, Siphoviridae, Unclassified
ENSG00000044574	<i>HSPA5</i>	Heat shock 70kDa proteins	56	Closteroviridae, Mimiviridae, Unclassified
ENSG00000063761	<i>ADCK1</i>		50	Pandoraviridae, Phycodnaviridae, Siphoviridae, Unclassified
ENSG00000164978	<i>NUDT2</i>	Nudix hydrolase family	46	Ascoviridae, Mimiviridae, Myoviridae, Unclassified
ENSG00000112031	<i>MTRF1L</i>		44	Myoviridae, Siphoviridae, Unclassified
ENSG00000124523	<i>SIRT5</i>	Sirtuins	42	Myoviridae, Podoviridae, Siphoviridae, Unclassified
ENSG00000143774	<i>GUK1</i>		38	Myoviridae, Pandoraviridae, Poxviridae, Siphoviridae, Unclassified
ENSG00000125871	<i>MGME1</i>	Exonucleases	37	Myoviridae, Siphoviridae, Unclassified
ENSG00000114956	<i>DGUOK</i>		34	Alloherpesviridae, Ascoviridae, Iridoviridae, Phycodnaviridae, Siphoviridae
ENSG00000166855	<i>CLPX</i>	AAA ATPases	32	Myoviridae, Podoviridae, Siphoviridae, Unclassified
ENSG00000103642	<i>LACTB</i>	Mitochondrial ribosomal proteins	32	Myoviridae, Siphoviridae
ENSG00000140521	<i>POLG</i>	DNA polymerases	32	Myoviridae, Podoviridae, Siphoviridae, Unclassified

Ensembl gene ID	Gene	HUGO gene family	Viral orthologue count	Viral families with orthologues
ENSG00000116863	<i>ADPRHL2</i>		31	Baculoviridae, Mimiviridae, Myoviridae, Phycodnaviridae, Unclassified
ENSG00000215021	<i>PHB2</i>		31	Myoviridae, Podoviridae, Siphoviridae, Unclassified
ENSG00000166199	<i>ALKBH3</i>	Alkylation repair homologs	30	Marseilleviridae, Mimiviridae, Pandoraviridae, Phycodnaviridae, Unclassified
ENSG00000141378	<i>PTRH2</i>	Cilia and flagella associated	30	Myoviridae, Poxviridae, Siphoviridae, Unclassified
ENSG00000113272	<i>THG1L</i>		30	Mimiviridae, Myoviridae, Siphoviridae, Unclassified
ENSG00000076043	<i>REXO2</i>	Exonucleases	28	Phycodnaviridae, Podoviridae, Poxviridae, Siphoviridae, Unclassified
ENSG00000205309	<i>NT5M</i>	5'-nucleotidases	27	Mimiviridae, Myoviridae, Siphoviridae, Unclassified
ENSG00000039650	<i>PNKP</i>	HAD Asp-based non-protein phosphatases	27	Mimiviridae, Myoviridae, Phycodnaviridae, Unclassified
ENSG00000166548	<i>TK2</i>		27	Herpesviridae, Hytrosaviridae, Iridoviridae, Myoviridae, Nudiviridae, Phycodnaviridae, Siphoviridae
ENSG00000173137	<i>ADCK5</i>		25	Phycodnaviridae, Unclassified
ENSG00000086061	<i>DNAJA1</i>	DNAJ (HSP40) heat shock proteins	25	Marseilleviridae, Mimiviridae, Phycodnaviridae, Unclassified
ENSG00000196365	<i>LONP1</i>	Serine proteases AAA ATPases	24	Mimiviridae, Phycodnaviridae, Unclassified
ENSG00000004779	<i>NDUFAB1</i>	NADH:ubiquinone oxidoreductase supernumerary subunits	24	Myoviridae, Unclassified
ENSG00000016391	<i>CHDH</i>		21	Mimiviridae, Unclassified
ENSG00000135821	<i>GLUL</i>		20	Mimiviridae, Unclassified
ENSG00000134905	<i>MCL1</i>	BCL2 family	20	Herpesviridae, Iridoviridae, Poxviridae
ENSG00000128050	<i>PAICS</i>	Purinosome	20	Myoviridae, Siphoviridae, Unclassified
ENSG00000134905	<i>CARS2</i>	Aminoacyl tRNA synthetases, Class I	19	Mimiviridae, Unclassified
ENSG00000242114	<i>MTFPI</i>		19	Mimiviridae, Phycodnaviridae
ENSG00000171552	<i>BCL2L1</i>	BCL2 family	18	Asfarviridae, Herpesviridae, Iridoviridae
ENSG00000105819	<i>PMPCB</i>	M16 metallopeptidases	18	Mimiviridae, Myoviridae
ENSG00000109971	<i>HSPA8</i>	Spliceosomal C complex	17	Closteroviridae, Mimiviridae, Phycodnaviridae, Unclassified
		Spliceosomal P complex		
		Heat shock 70kDa proteins		
ENSG00000197122	<i>SRC</i>	SH2 domain containing Src family tyrosine kinases	17	Retroviridae

Ensembl gene ID	Gene	HUGO gene family	Viral orthologue count	Viral families with orthologues
ENSG00000126215	<i>XRCC3</i>		17	Myoviridae, Unclassified
ENSG00000087074	<i>PPP1R15A</i>	Protein phosphatase 1 regulatory subunits	16	Ascoviridae, Asfarviridae, Baculoviridae, Herpesviridae, Hytrosaviridae, Poxviridae
ENSG00000074582	<i>BCS1L</i>	AAA ATPases Mitochondrial respiratory chain complex assembly factors	15	Myoviridae, Pandoraviridae, Siphoviridae, Unclassified
ENSG00000103018	<i>CYB5B</i>		15	Mimiviridae
ENSG00000189221	<i>MAOA</i>		14	Phycodnaviridae
ENSG00000100596	<i>SPTLC2</i>		13	Phycodnaviridae
ENSG00000106077	<i>ABHD11</i>	A hydrolase domain containing	12	Siphoviridae, Unclassified
ENSG00000244005	<i>NFS1</i>		12	Asfarviridae, Unclassified
ENSG00000104763	<i>ASAH1</i>		11	Mimiviridae
ENSG00000204389	<i>HSPA1A</i>	Heat shock 70kDa proteins TIM23 complex	11	Mimiviridae, Unclassified
ENSG00000144381	<i>HSPD1</i>	Chaperonins	11	Myoviridae, Podoviridae, Unclassified
ENSG00000103024	<i>NME3</i>	NME/NM23 family	11	Mimiviridae, Unclassified
ENSG00000126432	<i>PRDX5</i>	Peroxiredoxins	11	Myoviridae
ENSG00000100033	<i>PRODH</i>		11	Phycodnaviridae, Unclassified
ENSG00000178952	<i>TUFM</i>		11	Myoviridae, Siphoviridae, Unclassified
ENSG00000113013	<i>HSPA9</i>	Heat shock 70kDa proteins	10	Closteroviridae
ENSG00000137513	<i>NARS2</i>	Aminoacyl tRNA synthetases, Class II	10	Mimiviridae, Unclassified
ENSG00000139190	<i>VAMP1</i>	Vesicle associated membrane proteins	10	Phycodnaviridae
ENSG00000162129	<i>CLPB</i>	AAA ATPases Ankyrin repeat domain containing	9	Myoviridae, Unclassified
ENSG00000008405	<i>CRY1</i>		9	Mimiviridae, Phycodnaviridae, Unclassified
ENSG00000004864	<i>SLC25A13</i>	EF-hand domain containing Solute carriers	9	Mimiviridae
ENSG00000178700	<i>DHFRL1</i>		8	Myoviridae, Unclassified
ENSG00000162999	<i>DUSP19</i>	Atypical dual specificity phosphatases	8	Baculoviridae, Poxviridae, Siphoviridae
ENSG00000133315	<i>MACROD1</i>		8	Hesperidia, Iridoviridae, Gravidiae
ENSG00000133878	<i>DUSP26</i>	Atypical dual specificity phosphatases	7	Baculoviridae, Iridoviridae, Lipothrixviridae, Pandoraviridae, Unclassified
ENSG00000183597	<i>TANGO2</i>		7	Poxviridae, Unclassified
ENSG00000100823	<i>APEX1</i>		6	Mimiviridae, Phycodnaviridae, Unclassified
ENSG00000138363	<i>ATIC</i>	Purinosome	6	Myoviridae, Pandoraviridae
ENSG00000115204	<i>MPV17</i>		6	Phycodnaviridae
ENSG00000239672	<i>NME1</i>	NME/NM23 family	6	Mimiviridae, Unclassified
ENSG00000184428	<i>TOP1MT</i>	Topoisomerases	6	Mimiviridae, Unclassified
ENSG00000167468	<i>GPX4</i>	Selenoproteins	5	Poxviridae
ENSG00000115541	<i>HSPE1</i>	Chaperonins	5	Myoviridae
ENSG00000119912	<i>IDE</i>	M16 metallopeptidases	5	Mimiviridae, Unclassified

Ensembl gene ID	Gene	HUGO gene family	Viral orthologue count	Viral families with orthologues
ENSG00000143799	<i>PARP1</i>	Zinc finger PARP-type Poly(ADP-ribose) polymerases	5	Baculoviridae, Iridoviridae, Unclassified
ENSG00000150787	<i>PTS</i>		5	Myoviridae, Siphoviridae
ENSG00000167136	<i>ENDOG</i>		4	Unclassified
ENSG00000167085	<i>PHB</i>		4	Podoviridae, Siphoviridae, Unclassified
ENSG00000166532	<i>RIMKLB</i>		4	Myoviridae
ENSG00000087088	<i>BAX</i>	BCL2 family	4	Herpesviridae, Iridoviridae
ENSG00000171791	<i>BCL2</i>	BCL2 family	4	Asfarviridae, Herpesviridae
ENSG00000153291	<i>SLC25A27</i>	Solute carriers	4	Mimiviridae, Unclassified
ENSG00000174032	<i>SLC25A30</i>	Solute carriers	4	Mimiviridae
ENSG00000139644	<i>TMBIM6</i>	Transmembrane BAX inhibitor motif containing	4	Mimiviridae
ENSG00000188917	<i>TRMT2B</i>	tRNA methyltransferases	4	Mimiviridae, Unclassified
ENSG00000130985	<i>UBA1</i>	Ubiquitin like modifier activating enzymes	4	Mimiviridae, Unclassified
ENSG00000142208	<i>AKT1</i>	Cilia and flagella associated	3	Retroviridae
ENSG00000165644	<i>COMTD1</i>	Pleckstrin homology domain containing	3	Siphoviridae
ENSG00000166347	<i>CYB5A</i>	Seven-beta-strand methyltransferase motif containing	3	Phycodnaviridae, Unclassified
ENSG00000136628	<i>EPRS</i>	Aminoacyl tRNA synthetases, Class I	3	Unclassified
ENSG00000178445	<i>GLDC</i>	Aminoacyl tRNA synthetases, Class II	3	Myoviridae
ENSG00000105135	<i>ILVBL</i>		3	Phycodnaviridae, Unclassified
ENSG00000136003	<i>ISCU</i>		3	Siphoviridae
ENSG00000065427	<i>KARS</i>	Aminoacyl tRNA synthetases, Class II	3	Unclassified
ENSG00000247626	<i>MARS2</i>	Deafness associated genes	3	Unclassified
ENSG00000103202	<i>NME4</i>	Aminoacyl tRNA synthetases, Class I	3	Unclassified
ENSG00000138942	<i>RNF185</i>	NME/NM23 family	3	Mimiviridae
ENSG00000182199	<i>SHMT2</i>	Ring finger proteins	3	Iridoviridae
ENSG00000139131	<i>YARS2</i>	Aminoacyl tRNA synthetases, Class I	3	Myoviridae, Nudoviridae, Myoviridae, Unclassified
ENSG00000118520	<i>ARG1</i>		2	Phycodnaviridae, Unclassified
ENSG00000106105	<i>GARS</i>	Aminoacyl tRNA synthetases, Class II	2	Unclassified
ENSG00000171766	<i>GATM</i>		2	Myoviridae
ENSG00000165678	<i>GHITM</i>	Transmembrane BAX inhibitor motif containing	2	Phycodnaviridae
ENSG00000233276	<i>GPX1</i>	Selenoproteins	2	Poxviridae
ENSG00000112855	<i>HARS2</i>	Aminoacyl tRNA synthetases, Class II	2	Unclassified
ENSG00000135070	<i>ISCA1</i>		2	Phycodnaviridae
ENSG00000133703	<i>KRAS</i>	RAS type GTPase family	2	Retroviridae, Unclassified

Ensembl gene ID	Gene	HUGO gene family	Viral orthologue count	Viral families with orthologues
ENSG00000101247	<i>NDUFA5</i>	Seven-beta-strand methyltransferase motif containing Mitochondrial respiratory chain complex assembly factors	2	Siphoviridae, Unclassified
ENSG00000011052	<i>NME2</i>	NME/NM23 family	2	Alloherpesviridae, Unclassified
ENSG00000102078	<i>SLC25A14</i>	Solute carriers	2	Mimiviridae, Unclassified
ENSG00000164933	<i>SLC25A32</i>	Solute carriers	2	Nudiviridae
ENSG00000204394	<i>VARS</i>	Aminoacyl tRNA synthetases, Class I	2	Unclassified
ENSG00000154258	<i>ABCA9</i>	ATP binding cassette subfamily A	1	Poxviridae
ENSG00000135776	<i>ABCB10</i>	ATP binding cassette subfamily B	1	Siphoviridae
ENSG00000172955	<i>ADH6</i>	Alcohol dehydrogenases	1	Myoviridae
ENSG00000198610	<i>AKR1C4</i>	Aldo-keto reductases	1	Unclassified
ENSG000000081181	<i>ARG2</i>		1	Unclassified
ENSG00000130707	<i>ASS1</i>		1	Unclassified
ENSG00000103502	<i>CDIPT</i>		1	Unclassified
ENSG00000170312	<i>CDK1</i>	Cyclin dependent kinases	1	Herpesviridae
ENSG00000132423	<i>COQ3</i>	Seven-beta-strand methyltransferase motif containing	1	Phycodnaviridae
ENSG00000047230	<i>CTPS2</i>		1	Unclassified
ENSG00000087470	<i>DNM1L</i>		1	Unclassified
ENSG00000145982	<i>FARS2</i>	Aminoacyl tRNA synthetases, Class II	1	Unclassified
ENSG00000096384	<i>HSP90AB1</i>	Heat shock 90kDa proteins	1	Unclassified
ENSG00000087299	<i>L2HGDH</i>		1	Unclassified
ENSG00000002549	<i>LAP3</i>	Aminopeptidases	1	Unclassified
ENSG00000266472	<i>MRPS21</i>	Mitochondrial ribosomal proteins	1	Myoviridae
ENSG00000198804	<i>MT-CO1</i>	Mitochondrial complex IV: cytochrome c oxidase subunits	1	Unclassified
ENSG00000132781	<i>MUTYH</i>	DNA glycosylases	1	Unclassified
ENSG00000130414	<i>NDUFA10</i>	NADH:ubiquinone oxidoreductase supernumerary subunits	1	Podoviridae
ENSG00000128694	<i>OSGEPL1</i>		1	Siphoviridae
ENSG00000006757	<i>PNPLA4</i>	Patatin like phospholipase domain containing	1	Phycodnaviridae
ENSG00000135002	<i>RFK</i>		1	Unclassified
ENSG00000100316	<i>RPL3</i>	L ribosomal proteins	1	Retroviridae
ENSG00000073169	<i>SELO</i>		1	Myoviridae
ENSG00000183032	<i>SLC25A21</i>	Solute carriers	1	Mimiviridae
ENSG00000114120	<i>SLC25A36</i>	Solute carriers	1	Hytrosaviridae
ENSG00000162241	<i>SLC25A45</i>	Solute carriers	1	Unclassified
ENSG00000112096	<i>SOD2</i>		1	Unclassified
ENSG00000143374	<i>TARS2</i>	Aminoacyl tRNA synthetases, Class II	1	Unclassified
ENSG00000066651	<i>TRMT11</i>	tRNA methyltransferases	1	Pandoraviridae
ENSG00000100300	<i>TSPO</i>		1	Unclassified

Ensembl gene ID	Gene	HUGO gene family	Viral orthologue count	Viral families with orthologues
ENSG00000025708	<i>TYMP</i>	Minor histocompatibility antigens	1	Siphoviridae
ENSG00000177889	<i>UBE2N</i>	Ubiquitin conjugating enzymes E2	1	Unclassified

Appendix IV: Acetylation study species

Table 1. Vertebrate species included in the orthology dataset for the acetylation study.

Species	Group	Species	Group
<i>Homo sapiens</i>	Primates	<i>Eptesicus fuscus</i>	Bats
<i>Pan troglodytes</i>	Primates	<i>Myotis brandtii</i>	Bats
<i>Gorilla gorilla</i>	Primates	<i>Myotis lucifugus</i>	Bats
<i>Macaca fascicularis</i>	Primates	<i>Alligator mississippiensis</i>	Reptiles
<i>Callithrix jacchus</i>	Primates	<i>Alligator sinensis</i>	Reptiles
<i>Saimiri boliviensis</i>	Primates	<i>Crocodylus porosus</i>	Reptiles
<i>Carlito syrichta</i>	Primates	<i>Gavialis gangeticus</i>	Reptiles
<i>Otolemur garnettii</i>	Primates	<i>Thamnophis sirtalis</i>	Reptiles
<i>Oryctolagus cuniculus</i>	Primates	<i>Ophiophagus Hannah</i>	Reptiles
<i>Bos taurus</i>	Other Mammals	<i>Anolis carolinensis</i>	Reptiles
<i>Equus caballus</i>	Other Mammals	<i>Chelonia mydas</i>	Reptiles
<i>Canis familiaris</i>	Other Mammals	<i>Chrysemys picta</i>	Reptiles
<i>Orcinus orca</i>	Other Mammals	<i>Xenopus laevis</i>	Amphibia
<i>Loxodonta africana</i>	Other Mammals	<i>Gallus gallus</i>	Birds
<i>Dasyus novemcinctus</i>	Other Mammals	<i>Melopsittacus undulates</i>	Birds
<i>Sarcophilus harrisii</i>	Other Mammals	<i>Meleagris gallopavo</i>	Birds
<i>Monodelphis domestica</i>	Other Mammals	<i>Taeniopygia guttata</i>	Birds
<i>Ornithorhynchus anatinus</i>	Other Mammals	<i>Haliaeetus leucocephalus</i>	Birds
<i>Marmota marmota</i>	Rodents	<i>Falco peregrinus</i>	Birds
<i>Ictidomys tridecemlineatus</i>	Rodents	<i>Corvus brachyrhynchos</i>	Birds
<i>Heterocephalus glaber</i>	Rodents	<i>Nipponia nippon</i>	Birds
<i>Cavia porcellus</i>	Rodents	<i>Egretta garzetta</i>	Birds
<i>Chinchilla lanigera</i>	Rodents	<i>Charadrius vociferous</i>	Birds
<i>Octodon degus</i>	Rodents	<i>Cuculus canorus</i>	Birds
<i>Castor canadensis</i>	Rodents	<i>Calypte anna</i>	Birds
<i>Dipodomys ordii</i>	Rodents	<i>Chaetura pelagica</i>	Birds
<i>Jaculus jaculus</i>	Rodents	<i>Columba livia</i>	Birds
<i>Microtus ochrogaster</i>	Rodents	<i>Anas platyrhynchos</i>	Birds
<i>Mesocricetus auratus</i>	Rodents	<i>Struthio camelus</i>	Birds
<i>Peromyscus maniculatus</i>	Rodents	<i>Latimeria chalumnae</i>	Fish
<i>Neotoma lepida</i>	Rodents	<i>Oryzias latipes</i>	Fish
<i>Mus musculus</i>	Rodents	<i>Oreochromis niloticus</i>	Fish
<i>Rattus norvegicus</i>	Rodents	<i>Esox Lucius</i>	Fish
<i>Pteropus alecto</i>	Bats	<i>Danio rerio</i>	Fish
<i>Pteropus vampyrus</i>	Bats	<i>Lepisosteus oculatus</i>	Fish
<i>Rousettus aegyptiacus</i>	Bats	<i>Callorhinchus milii</i>	Fish